Attribute Weighted Fuzzy Interpolative Reasoning

Fangyi Li

Supervisors: Prof. Qiang Shen
Dr. Changjing Shang

Ph.D. Thesis
Department of Computer Science
Faculty of Business and Physical Sciences
Aberystwyth University

April 20, 2020

Declaration and Statement

DECLARATION
This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

STATEMENT 1
This thesis is the result of my own investigations, except where otherwise stated. Where **correction services**[1] have been used, the extent and nature of the correction is clearly marked in a footnote(s).
Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... (candidate)

Date ....

STATEMENT 2
I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

[1]This refers to the extent to which the text has been corrected by others.

Abstract

Approximate reasoning systems facilitate fuzzy inference through manipulating fuzzy if-then rules. Fuzzy rule interpolation (FRI) supports such reasoning with sparse rule bases where certain observations may not match any existing fuzzy rules. While offering a potentially powerful inference mechanism, in the current literature, it is typically assumed that all antecedent attributes in the rules are of equal significance in deriving the consequents. This is a strong assumption in practical applications, thereby, often leading to less accurate interpolated results. Recently, interesting techniques have been reported for achieving weighted interpolative reasoning. However, they either employ attribute weights that are obtained using additional information (rather than just the given rules) or fail to enable the individual attribute weights to be integrated systematically with the corresponding FRI procedures. To devise a weighted interpolative reasoning that works effectively and efficiently, two major concerns need to be addressed. First, how the rule antecedent weights can be generated automatically and efficiently, without requiring further observations or triggering the entire unweighted FRI system. Second, how the generated weights may be integrated within any unweighted FRI mechanism. A further associated issue is how a weighted FRI method may be transplanted to another underlying FRI where no individual attribute weight is involved once the weights of rule antecedents are available.

This thesis proposes a weighted fuzzy interpolative reasoning mechanism, leading to novel FRI approaches that significantly reinforce the power of approximate reasoning. It works by exploiting attribute ranking techniques to help determine the relative importance of rule antecedent attributes involved in a sparse rule base. In particular, the proposed approach employs feature selection (FS) techniques to adjudge the relative significance of individual attributes and therefore, in order to differentiate the contributions of the rule antecedents and their impacts on FRI. This is feasible because FS provides a readily adaptable mechanism for evaluating and ranking attributes, being capable of selecting more informative features. Without requiring any acquisition of real observations, based on the originally given sparse rule base, the individual weights are computed using a set of training samples that are artificially created from the rule base through an innovative *reverse engineering* procedure. This weight generation procedure is general as it allows for any established

ranking method to be utilised to score the attributes without adversely affecting the interpolative inference accuracy. Given the generated weights of rule antecedent attributes, this thesis further presents three FRI approaches, each based on different type of fuzzy interpolative reasoning technique, for systematically integrating the weights within the FRI procedure. Such a weighted approach integrates the learned weights explicitly with all computational steps of the interpolation process. The implementation of each weighted FRI mechanism is of generality as it is achieved independently of the weight generation method. Thus, the underlying generic techniques can be extended to supporting any other FRI which involves multiple rule antecedents which are not assigned with individual weights.

The proposed weighted FRI approaches have been statistically evaluated through a range of experimentations against various benchmark datasets. The results are reported in this thesis, demonstrating the superior and robust performance of the weighted methods over their originals (where the rule antecedent attributes are of equal significance). A specific and important outcome that is supported by attribute ranking is that only two (i.e., the least number of) nearest neighbouring rules are required to perform accurate interpolative reasoning. This avoids the need of both searching for and computing with multiple rules beyond the immediate neighborhood of a given observation, thereby significantly enhancing computational efficiency. The proposed weight generation and weighted FRI mechanisms are integrated with the standard compositional rule of inference to develop application systems to perform real-world pattern recognition tasks, including classification and prediction (which in turn, involves both multivariate regression and time series prediction). Particularly, the thesis reports on a novel fuzzy rule-based diagnostic system for mammographic mass classification. This system is able not only to derive a conclusion for unknown observed masses that have no rules to match, but also to produce the diagnostic outcomes that are interpretable, thanks to the semantics-rich fuzzy rules with attribute values represented in linguistic terms. The success in all such realistic applications demonstrates the practicality of the proposed techniques for attribute weighted fuzzy interpolative reasoning.

Acknowledgements

Many thanks to my first supervisor, Prof. Qiang Shen, for his invaluable supervision and stimulating encouragement throughout this project. Without him, this PhD and thesis would never have happened. Thanks must also go to Dr. Changjing Shang, my second supervisor, for her patient guidance during the research. In addition, many thanks go to all fellow researchers in the Advanced Reasoning Group, especially Dr. Neil Mac Parthaláin for his warm-hearted discussion and helpful advice, and my good friends (e.g., Jing Yang and Pu Zhang) for their encouragement and support. Finally, my sincere gratitude to my beloved parents and Hang for their love and the unwavering support through all the past four years.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

FUZZY set theory [Zadeh, 1965], and its extension, fuzzy logic, have gained rapid developments in a variety of scientific areas, including mathematics, engineering, and computer science. They also have been successfully applied for many real-world problems [Ross, 2005, Terano et al., 2014, Zimmermann, 2011, Li et al., 2016], such as systems control, fault diagnosis and computer vision, as an effective tool to address the issues of imprecision and vagueness in modelling and reasoning. This makes systems developed on the basis of fuzzy sets and fuzzy logic a core paradigm in the field of soft computing [Bonissone, 1997, Zadeh, 1994], forming sharp contrast with the conventional hard computing systems based on boolean logic and numerical analysis. In particular, fuzzy expert systems exploit the tolerance for imprecision, partial truth and approximations to achieve close resemblance with human activity and reasoning intuition. Many of which have been developed using the idea of approximate reasoning (also known as linguistic reasoning), reflecting the manner of human cogitation and leading to new, more human interpretable, intelligent systems.

1.1 Approximate Inference

Reasoning with imprecise information is one of the central topics of fuzzy logic. In general, an approximate reasoning system can be formalized as a fuzzy *if-then* rule-based inference mechanism that derives a conclusion given an input observation. It consists of linguistic variables, fuzzy rules and a fuzzy inference mechanism. Linguistic variables facilitate the interpretation of linguistic expressions in terms of fuzzy quantities of certain underlying mathematical semantics. Fuzzy inference rules are a set of rules that associate input and output data of an underlying system, to model either historical data acquired from the system or expert opinions regarding the system, or a mixture of both, typically expressed in linguistic terms. Based on such rules, a fuzzy inference mechanism is encoded to implement the process of approximate reasoning, through manipulation among the fuzzy inference rules in response to any new input data.

Various techniques have been established to build fuzzy systems. In particular, many have been seen to implement generalised *modus ponens* that facilitates reasoning when provided with imprecise inputs, mostly by following the basic idea of Compositional Rule of Inference (CRI) [Zadeh, 1973]. That is, a fuzzy rule is regarded as a triplet that consists of an antecedent, and a consequent, both of which are linguistic variables and linked through a fuzzy relation. The CRI performs its work given a fuzzy rule or a fuzzy rule base (i.e., a finite collection of fuzzy rules) in the following manner: *if the system input coincides with the antecedent of a fuzzy rule, then the output should coincide with the consequent that corresponds to the antecedent of that fuzzy rule* [Fuller, 2000]. This reflects the property of CRI inference being a generalisation of the *modus ponens* of classical logic, reflecting the intuition of similar inputs normally deriving similar outputs.

The law of CRI has been successfully applied in approximate inference and fuzzy control, for example, the Mamdani's fuzzy logic controller [Mamdani and Assilian, 1999] was implemented this way. However, CRI is unable to draw a conclusion when a rule base is not dense but sparse. Sparse, or incomplete, rule bases considered here are not referring to the quantity of rules in a given rule base, but to the coverage of the problem domain by the antecedents of rules regarding the universe of discourse. That is, an input observation may have no overlap with any of the rules available and hence, no rule may be executed to derive the required consequent by directly applying CRI.

1.2 Fuzzy Interpolative Reasoning

Fuzzy rule interpolation (FRI) facilitates approximate reasoning in fuzzy rule-based systems when only sparse knowledge is available [Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b]. It addresses the key limitation of conventional fuzzy rule-based systems that work using CRI [Zadeh, 1973], requiring a dense fuzzy rule base which fully covers the entire problem domain. If, however, the problem domain is not completely covered by the given rules, there may exist observations that fail to activate any existing rules to compute a required inference outcome. Resolving real-world problems frequently involves the use of such sparse rule bases, since a dense rule base is especially impracticable in a multidimensional environment where the number of rules increases exponentially as the input variables and the fuzzy linguistic labels associated with each variable increase. FRI plays an approximate and useful role in such situations explicitly where only an incomplete rule base is available. It works with this form of sparse knowledge, attempting to reduce, if not to completely remove, the restriction of CRI for cases where no conclusion may be derived due to no rules matching a new observation. This offers an alternative way to infer an approximately interpolated outcome, accomplishing the so-called fuzzy interpolative reasoning.

FRI essentially makes two contributions to the development of fuzzy rule-based systems. It not only facilitates the assistance of reasoning on sparse rule bases [Burkhardt and Bonissone, 1992], but also offers the potential for a reverse application where the rule base may be so dense that model simplification is required. That is, FRI can be utilised to simplify the complexity of fuzzy rule bases through say, a procedure of iteratively replacing two existing rules with an interpolated one [Koczy and Hirota, 1997], thereby eliminating those fuzzy rules which may be approximated from their neighbouring ones. Whilst both contributions are based on manipulating certain defined neighbouring rules, the latter is beyond the scope of this thesis research. That is, this work is focussed on performing inference with a sparse rule base. In supporting fuzzy interpolative inference on sparse rule bases, the concept of neighbouring fuzzy rules is indeed fundamental, of which the rule antecedent parts have the highest similarity with the given observation. Note that the goal of FRI is not to produce an interpolated rule through interpolative reasoning, but to compute an interpolated consequent that corresponds to the input observation. In so doing, FRI achieves the inference task with respect to the observations that originally have no conclusions to be drawn due to the sparseness of the fuzzy rule base.

As an inference mechanism, FRI starts to reach its goal from the selection of the nearest neighbouring (aka. the closest) rules in the fuzzy sparse rule base with regards the given unmatched observation. Such chosen rules form the basis for conducting fuzzy interpolation. Two major categories have been seen in the literature to implement fuzzy interpolative reasoning: 1) the α-cut based interpolations (e.g., [Chang et al., 2008, Kóczy and Hirota, 1993a, Tikk and Baranyi, 2000, Yam et al., 2006]) and 2) the intermediate rule based interpolations (e.g., [Baranyi et al., 2004, Chen et al., 2016, Huang and Shen, 2006, Jin et al., 2014, Yang et al., 2017]). This categorisation depends upon whether the computation of the interpolated result is accomplished through a process of construction and transformation of an intermediate rule first. As such, FRI methods may also be organised in two groups, respectively termed as non-transformation based and transformation based FRI [Chen and Adam, 2018]. The pioneering work for fuzzy interpolative reasoning, as of the techniques reported in [Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b] and their extensions, form the most typical non-transformation based FRI. For those relying on transforming intermediate rules, a family of scale and move transformation-based FRI (termed as T-FRI), such as those given in [Huang and Shen, 2006, Huang and Shen, 2008, Jin et al., 2014, Yang et al., 2017], have been popularly studied and widely applied.

1.3 Two Key Issues Concerned within Thesis

In resolving practical real-world problems, multidimensional input variables are a common issue. Fortunately, many FRI methods exist in the literature that are capable of dealing with interpolation, by the use of fuzzy rules that involve multiple conditional (interchangeably termed antecedent hereafter) variables. Nonetheless, there is a common problem existing in these FRI approaches, where the conditional attributes within the rules are presumed to be of equal significance for interpolation. Thus, inaccurate and even incorrect interpolated outcomes may result since different domain attributes may generally make different contributions to the decision making process.

Recently however, a number of methods have been proposed for FRI working on multiple conditionals associated with different weights (e.g., [Chen and Chang, 2011b, Chen and Chen, 2016, Chen et al., 2009, Cheng et al., 2015, Diao et al., 2014]). Nevertheless, two key questions remain to be further investigated in developing such weighted FRI:

1. how the weights are generated; and
2. whether and how the weights are integrated within the underlying, non-weighted FRI.

Regarding the first question, one of the possible answers is to assign the predefined weights by domain experts [Li et al., 2005]. However, this will require human intervention and hence adversely reduce the flexibility and automation level of the resulting fuzzy systems. Automatic weight learning schemes are obviously preferred. In particular, work already exists that utilises Genetic Algorithms (GA) to induce from data the weights of rule conditionals, thereby strengthening the effectiveness of FRI [Chen and Chang, 2011b]. Yet, such techniques introduce much more additional computation and also the required specification of many GA parameters. Alternatively, the weights may be determined by a distance measure between the information content of an observation and that of an conditional attribute within a given rule. The information concerned is related to the characteristic points of the fuzzy sets that specify the corresponding attributes, including, for example, the central point [Chen and Chen, 2016] or the ranking value [Cheng et al., 2015] of a fuzzy set. The weights are then assigned differently to the same conditional fuzzy set that appears in each and every different rule, incurring significant extra computation and reducing the interpretability of the weighted rules. Different from these, the weight learning schemes as reported in [Chen et al., 2009], form an implementation of the "wrapper" approach, thereby mixing up FRI-based inference and learning from data.

The second question arises due to the observation that the existing techniques typically work by artificially creating a simple overall weight to each of the rules before the weighted rules are run in FRI. Such weights are normally computed through aggregating the weights calculated for individual conditionals, thereby requiring additional aggregation procedures. Situations generally become even more complicated if different fuzzy interpolative reasoning systems are considered to exploit the weighted rules to perform FRI, assuming the use of different supporting techniques (e.g., piecewise fuzzy entropies [Chen and Chen, 2016] and ranking scores [Cheng et al., 2015] of the fuzzy values involved in the rules). The resulting weights may be exploited rather differently, depending on what underlying FRI mechanism is employed. Most significantly, within existing methods, the computed weights are decoupled from the internal working procedures which utilise them.

This makes the interpretation of the resulting FRI process and hence, that of the interpolated results more difficult than the interpretation of the results from explicitly integrating the weights and individual FRI procedures.

From the above, it is therefore desirable to conduct other research into the aforementioned two issues for developing more effective and efficient FRI techniques. This thesis puts forward such techniques to generate weights for individual rule antecedents and to facilitate weighted fuzzy interpolative reasoning.

1.4 Feature Evaluation

There have been a number of proposals made in attempting to resolve the first issue (that is *how the weights can be generated*), for assessing the capabilities of domain attributes in terms of their influence upon the potential consequent that depends on these attributes. In particular, the techniques of feature selection (FS) provide an effective measures to facilitate such a solution. This is because FS [Dash and Liu, 1997, Liu and Motoda, 2012] aims to discover a minimal subset of features that are most predictive of a given outcome. It generally follows a four-step procedure: generation, evaluation, termination and validation. Feature subsets are generated via a certain search procedure amongst the family of subsets of the original feature set. These feature subsets are then evaluated individually with regard to a given quality measure. The process of searching for a reduced feature subset is terminated if the measured quality degree reaches a satisfactory level. Finally, a selected feature subset is validated with respect to the application problem at hand.

In developing effective FS mechanisms, much work has been carried out regarding the second step that evaluates the quality of a candidate feature subset [Cui et al., 2010, Dash and Liu, 2003, Jensen and Shen, 2009, Zeng and Cheung, 2010], including those directly assessing and ranking individual features [He et al., 2006, Kononenko, 1994, Uğuz, 2011]. For any reasoning system (be it fuzzy or boolean), different ranking scores of features or domain attributes imply different contributions of them to the inference outcome. Inspired by this observation, feature evaluation methods may be adapted to score the significance of individual rule antecedents, for generating the corresponding individual weights. This methodology is to be followed in the work of this thesis.

1.5 Proposal for Attribute Weighted Fuzzy Interpolative Reasoning

This thesis proposes a novel fuzzy rule-based interpolative reasoning mechanism that is guided by the attribute weights to address the two questions raised earlier.

Firstly, the attribute weighting scheme is enabled by an innovative *Reverse Engineering* procedure, which reduces the sparsity of the given rule base by generating an artificial training decision table from the given sparse rule base. The essential idea is to reformulate all rules in the rule base into a common representation, where each (possibly) missing value of any rule antecedent is replaced by one of the alternative fuzzy values from its domain. All these reformulated rules, artificial or original, are collated for evaluation of the relative significance degrees of the individual attributes. The weights of the attributes are individually measured using a certain feature ranking method, which is implemented by modifying the feature evaluation procedure extracted from a selected FS technique. Different types of FS method may be adopted for such use without significantly affecting the level of performance improvement over the conventional unweighted FRI. In so doing, it is expected that resulting weighted FRI approach will offer flexibility in its implementation.

Secondly, to minimise the adverse effect of existing FRI methods that is caused by assuming all attributes having equal significance, weights are introduced to rule antecedent attributes. In particular, for weighted T-FRI, individual attribute weights are integrated with every procedure of the underlying unweighted algorithm. In this work, T-FRI is used as a representative in implementation, unless otherwise stated. Hence, there will be three procedures that involve such integration: the selection of the nearest neighbouring rules, the construction of intermediate rules, and the computation of scale and move transformation factors. All computational steps in the original T-FRI, which effectively deals with evenly calculated average of the attribute values, will be improved by a weighted aggregation of the corresponding components. This weighting scheme over unweighted (interchangeably termed non-weighted hereafter) FRI will also be extended to two other popular FRI approaches, as of [Kóczy and Hirota, 1993a] and [Chang et al., 2008], demonstrating the generalisation capability of integrating attribute weights within unweighted fuzzy interpolative reasoning.

With such extended work, it is important to verify whether the resulting algorithms outperform their originals and if so, to what extent and effect. One particular aspect of this thesis is therefore to show that supported by attribute ranking, only two (i.e., the least number of) nearest adjacent rules are required to perform accurate interpolative reasoning. This helps increase computational efficiency, without the need of searching for and operating on multiple rules beyond the immediate neighborhood of a given observation.

Finally, also for the purpose of verifying and demonstrating the potential of the proposed attribute weighted FRI, it will be systematically applied to facilitate fuzzy rule-based interpolative reasoning to perform classification and prediction tasks. It will also be adapted for accomplishing real-world mammographic mass image analysis in support of medical diagnosis.

1.6 Thesis Structure

The rest of this thesis is structured as follows, with an illustrative reading guide given at the end of this chapter. In particular, an indication of directly relevant, peer-reviewed publications produced as a result of this research is shown, where journal articles are denoted by Jx and conference papers by Cy, with x and y indexing the journal and conference paper number, respectively.

Chapter 2: Background

This chapter reviews the preliminary knowledge that is relevant to this project, including an overview of the seminal FRI approaches and the existing weighted fuzzy interpolative reasoning techniques, and an outline of different types of attribute evaluation method that are extracted from FS techniques. The contents of part of the literature review in this chapter are currently under review (J1) for journal publication.

Chapter 3: Weight Learning from Rule Bases

This chapter proposes a novel weight induction mechanism that learns the weights of rule antecedent attributes given a (sparse) fuzzy rule base only. It is implemented by the use of an innovative *Reverse Engineering* procedure for generating the training data from the given rule base, and by the adaptation of attribute evaluation approaches for producing the required individual weights. A practical illustrative case study is also introduced in this chapter, which is utilised to demonstrate the working procedure of the entire weight learning process. The contents of this chapter have been published in C1 [Li et al., 2017b], J2 [Li et al., 2018c] and J3 [Li et al., 2018a].

Chapter 4: Weighted Transformation-based FRI

This chapter presents the framework for weighted fuzzy interpolative reasoning, provided with learned rule antecedent weights (following the work of Chapter 3). In particular, the work is implemented by adapting the popular scale and move transformation-based FRI (T-FRI) [Huang and Shen, 2006, Huang and Shen, 2008] (that only deals with rules whose antecedent attributes are of equal significance), with the weights being employed to modify both the computation process of closest rule selection and that of rule interpolation. In addition, this chapter recalls and continues the illustrative case study where the attribute weights have been generated in the last chapter, to explain how the weighted T-FRI performs its work, completing the illustration for the entire process of the proposed weighted fuzzy interpolative reasoning. Part of this chapter has been published in C2 [Li et al., 2017a], J2 [Li et al., 2018c] and J3 [Li et al., 2018a].

Chapter 5: Evaluation of Attribute Weighted Transformation-based FRI

This chapter evaluates the weighted fuzzy interpolative reasoning framework as proposed in Chapter 3 and Chapter 4. The computational complexity of the framework is first analysed from the theoretical viewpoint. It is followed by an experimental evaluation over two realistic pattern recognition tasks, namely classification and prediction over a range of benchmark datasets. Comparative studies demonstrate that the proposed research helps minimise the adverse impact of the equal significance assumption made in the conventional FRI techniques, significantly improving the accuracy of interpolated results. The contents of this chapter have been published in J2 [Li et al., 2018c] and J4 [Li et al., 2019a].

Chapter 6: Extensions to Alternative FRI Approaches

This chapter presents a further development of weighted FRI to enhance two other commonly used FRI algorithms (namely, those first presented in [Kóczy and Hirota, 1993a] and [Chang et al., 2008]), by following the ideas of weighted T-FRI (as presented in Chapter 4). The improvement of classification accuracies is highlighted by systematic comparisons. Importantly, it is shown that the best performance is achieved when the number of the nearest neighbouring rules required to perform weighted FRI is indeed the smallest, i.e., two. Part of this chapter has been published in J5 [Li et al., 2019c].

Chapter 7: Application of Weighted Fuzzy Interpolative Reasoning to Interpretable Mammographic Mass Classification

This chapter presents an application of the weighted fuzzy interpolative reasoning mechanism to a significant real-world problem, for addressing mammographic mass classification in support of breast cancer risk analysis. The implemented system is able to derive a conclusion for unknown observed masses that have no rules to match. The results show that, apart from achieving accurate classification, the diagnostic outcomes are interpretable. This latter aspect is due to the fact that the rules are learned from selected features in terms of mass geometric and density properties, with the feature values also represented in linguistic terms. The contents of this chapter have been published in C3 [Li et al., 2018b] and J6 [Li et al., 2019b].

Chapter 8: Conclusion

The thesis is concluded and challenging further work discussed in this chapter. In particular, promising research as reported in C4, J7 and J8 is identified as the candidates to be integrated with the attribute weight learning mechanism to further strengthen their potential.

Appendices

Appendix A outlines a data-driven iterative rule base generation procedure that to produce fuzzy rules are used in the experimental evaluations of Chapters 5, 6 and 7. Appendix B summarises the abbreviations employed throughout the thesis.

In short, Fig. 1.1 provides a guidance for assisting the read of this thesis (bar Appendices). In addition, this figure also indicates the basic relationship between the thesis work and the resulting publications, including 7 journal articles and 4 conference papers already published and one currently under review for journal publication. More details of the publications are summarised in Tables 1.1 and 1.2.

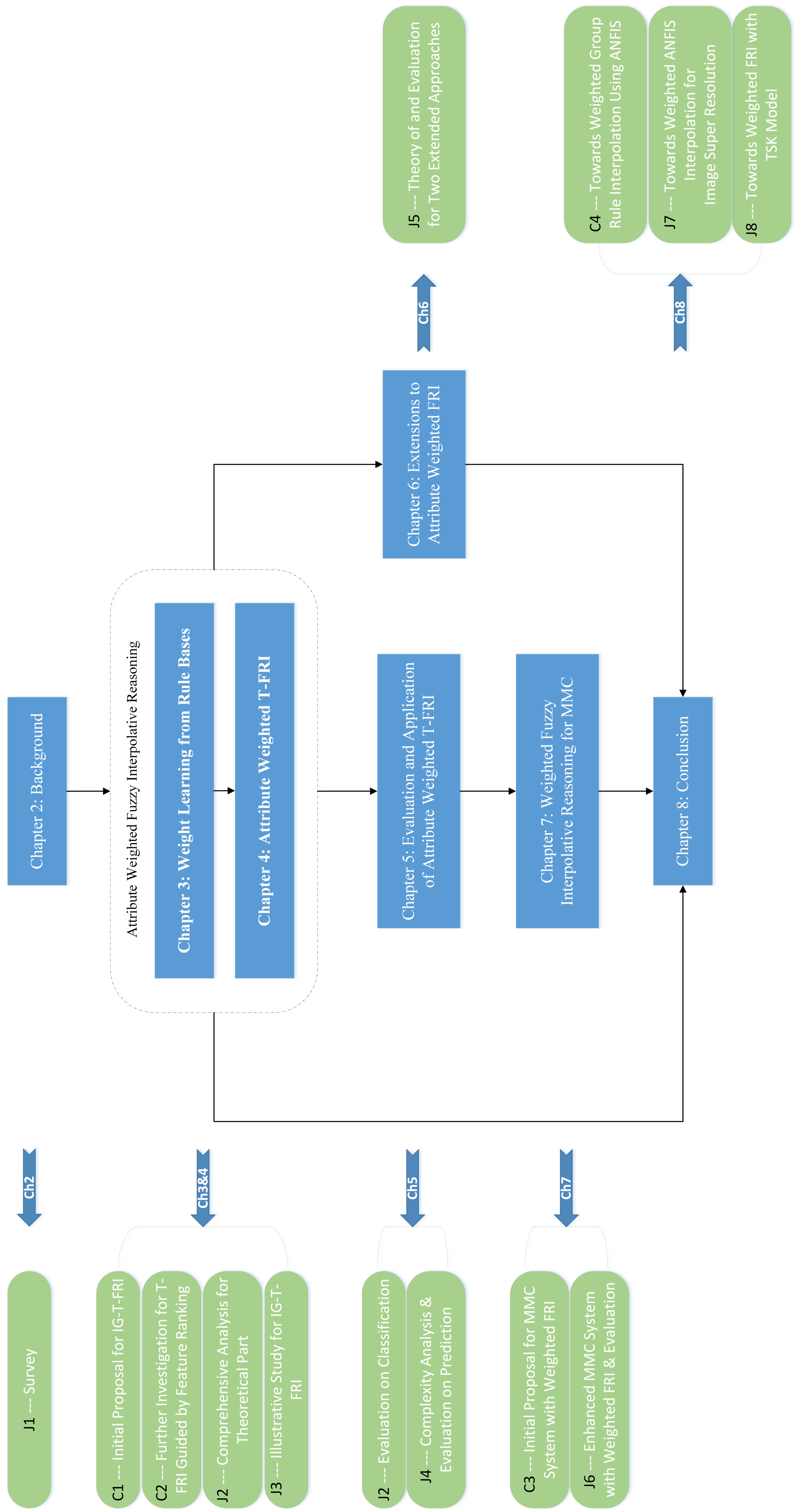


Figure 1.1: Guidance for thesis reading with links between chapters and publications.

Table 1.1: Publications Arising from the Thesis — Journal Articles

No.	Publication
J1	"Approximate reasoning with fuzzy rule interpolation: Background and recent advances". **Fangyi Li**, Changjing Shang, Ying Li, Jing Yang and Qiang Shen. Under review for journal publication.
J2	"Fuzzy Rule Based Interpolative Reasoning Supported by Attribute Ranking". **Fangyi Li**, Changjing Shang, Ying Li, Jing Yang and Qiang Shen. *IEEE Transactions on Fuzzy Systems* 26, no. 5 (2018): 2758-2773.
J3	"Improving fuzzy rule interpolation performance with information gain-guided antecedent weighting". **Fangyi Li**, Ying Li, Changjing Shang and Qiang Shen. *Soft Computing* 22, no. 10 (2018): 3125-3139.
J4	"Fuzzy Knowledge-Based Prediction Through Weighted Rule Interpolation". **Fangyi Li**, Ying Li, Changjing Shang and Qiang Shen. *IEEE Transactions on Cybernetics* (Early Access) (2019). DOI: 10.1109/TCYB.2018.2887340.
J5	"Interpolation with Just Two Nearest Neighbouring Weighted Fuzzy Rules". **Fangyi Li**, Changjing Shang, Ying Li, Jing Yang and Qiang Shen. *IEEE Transactions on Fuzzy Systems* (Early Access) (2019). DOI: 10.1109/TFUZZ.2019.2928496.
J6	"Interpretable Mammographic Mass Classification with Fuzzy Interpolative Reasoning". **Fangyi Li**, Changjing Shang, Ying Li and Qiang Shen. *Knowledge-Based Systems* (2019): 105279.
J7	"ANFIS construction with sparse data via group rule interpolation". Jing Yang, Changjing Shang, Ying Li, **Fangyi Li** and Qiang Shen. *IEEE Transactions on Cybernetics* (Early Access) (2019). DOI: 10.1109/TCYB.2019.2952267.
J8	"A new approach for transformation-based fuzzy rule interpolation". Tianhua Chen, Changjing Shang, Jing Yang, **Fangyi Li** and Qiang Shen. *IEEE Transactions on Fuzzy Systems* (Early Access) (2019). DOI: 10.1109/TFUZZ.2019.2949767.

Table 1.2: Publications Arising from the Thesis — Conference Papers

No.	Publication
C1	“Guiding fuzzy rule interpolation with information gains”. **Fangyi Li**, Changjing Shang, Ying Li and Qiang Shen. *In Advances in Computational Intelligence Systems*, pp. 165-183. Springer, Cham, 2017. (Best student paper award at the 17th Annual Workshop on Computational Intelligence)
C2	“Feature ranking-guided fuzzy rule interpolation”. **Fangyi Li**, Changjing Shang, Ying Li and Qiang Shen. *In 2017 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE), pp. 1-6. IEEE, 2017.
C3	“Feature Ranking-Guided Fuzzy Rule Interpolation for Mammographic Mass Shape Classification”. **Fangyi Li**, Changjing Shang, Ying Li and Qiang Shen. *In 2018 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE), pp. 1-7. IEEE, 2018.
C4	“Generating ANFISs through rule interpolation: An initial investigation”. Jing Yang, Changjing Shang, Ying Li, **Fangyi Li** and Qiang Shen. *In UK Workshop on Computational Intelligence*, pp. 150–162, Springer, 2018.

Chapter 2

Background

FUZZY interpolative reasoning plays an important role in fuzzy rule based inference systems, facilitating the extension of the capability of approximate reasoning when dealing with incomplete knowledge. This is because fuzzy rule interpolation (FRI) is able to produce an approximate interpolated outcome by the use of limited fuzzy rules that fail to derive a conclusion to an unmatched input observation. FRI techniques have been continuously investigated for decades, resulting in various types of approach. Recent studies have shown great interesting in developing an enhanced FRI where the rule antecedent attributes are associated with relative weights, signifying their different importance in influencing the generation of the conclusion, thereby improving the interpolative inference performance. Research from this viewpoint essentially opens a wide field, including the inevitable point of attribute evaluation and weight generation in particular.

In this chapter, the preliminary background knowledge closely relevant to the developments of the subsequent chapters is reviewed. This ranges from the fundamental FRI techniques, the present weighted fuzzy interpolative reasoning mechanisms, and the underlying feature evaluation approaches that will be adopted for assessing the rule antecedent attributes. To facilitate the demonstration of the relevant methodology, basic notations that will be used throughout the thesis are described first.

2.1 Basic Notations for Fuzzy Rule-based Inference Systems

In general, a fuzzy rule-based system has as its key component a set of *if-then* rules, each of which takes fuzzy or crisp terms that represent specifications of the input variables and associates these with the output of a certain problem description. In general, a rule may involve multiple output attributes as well as multiple input variables, but a multiple output rule can always be equivalently expressed by several single output rules. Without losing generality, only rules which have a single output class are considered in this work.

Formally, a typical fuzzy rule model essentially contains two key elements $\langle R, Y \rangle$ in describing a given problem: A non-empty finite set of domain attributes $Y = A \cup \{z\}$, where $A = \{a_j | j = 1, 2, \ldots, m\}$ represents the set of input antecedent attributes and z stands for the consequent, and a non-empty finite set of fuzzy rules $R = \{r^1, r^2, \ldots, r^N\}$. In many conventional fuzzy rule-based systems, including systems implemented with FRI techniques, a given rule $r^i \in R$ and an observation o^* are often expressed generally as follows:

$$\begin{aligned} r^i \ &: if\ a_1\ is\ A_1^i\ and\ a_2\ is\ A_2^i\ and\ \cdots\ and\ a_m\ is\ A_m^i, \\ &\quad then\ z\ is\ B^i \\ o^* \ &: a_1\ is\ A_1^*\ and\ a_2\ is\ A_2^*\ and\ \cdots\ and\ a_m\ is\ A_m^* \end{aligned} \tag{2.1}$$

where A_j^i and A_j^* denote the fuzzy set values taken by the antecedent attribute a_j in r^i and o^*, respectively; and B^i represents the fuzzy set value of the consequent attribute z in r^i.

Fuzzy values of both the rule antecedents and the consequent are in general represented by fuzzy sets. The concept of fuzzy sets was introduced by L. Zadeh [Zadeh, 1965]. Informally, the definition of a fuzzy set given by L. Zadeh can be stated as follows: A fuzzy set is a class with a continuum of membership grades. Thus, a fuzzy set A in a universe of discourse X is characterised by a membership function (MF) A which associates each element $x \in X$ with a real number $A(x) \in [0, 1]$. This is interpreted such that $A(x)$ is the membership grade of x belonging to the fuzzy set A [Bede, 2013].

As can be seen from the above, the MF $A : X \to [0, 1]$ distinguishes fuzzy sets from classical boolean sets. Unlike a classical set with clear boundaries, i.e., $x \in A$ or $x \notin A$, which excludes any other possibility, the property of the membership function enables fuzzy sets to model partial degrees to which a variable or attribute is deemed to take a certain underlying real or categorical value. Such fuzzy sets are often assigned with linguistic terms to help capture and reflect human interpretation of imprecise measurements or descriptions.

Particularly, when the universe of discourse X consists of the line of real number $\mathbb{R}$, any type of continuous functions can be used as an MF, provided that a set of parameters is given to specify the appropriate meanings of the MF. In this case, it is impractical to list all the pairs defining an MF, even if imposing the constraint that all MFs are convex in topology to ease the expression of common sense interpretation of belongingness. Fortunately, only a small number of types of MF that are typically used in practice. Basically, there are two main categories in terms of their properties: smoothness and linearity, which are: i) polygonal (piecewise linear) fuzzy sets, including triangular shaped, trapezoidal shaped MF, etc., and ii) nonlinear fuzzy sets, typically including Gaussian, Generalised bell-shaped, and Sigmoid MFs.

Polygonal fuzzy sets are generally represented by their characteristic points (CPs) in ascending order (which are defined mathematically as the odd points of the membership function [Huang and Shen, 2008]), and nonlinear ones by the defining parameters that are used to specify each nonlinear function. The choice of different MFs relies on the specific requirements of a given application. Amongst the family of all possible functions, triangular MFs and trapezoidal MFs have been used extensively, especially for real-time implementations, thanks to their simple representation and computational efficiency. In developing FRI methods, different MFs have been exploited to implement various approaches. In particular, procedures employing triangular MFs and/or trapezoidal MFs can be seen as specific cases of those which utilise more complex polygonal fuzzy sets. It is difficult to have a generic closed form representation that unifies all FRI processes as they are dependent upon the MFs used. Nevertheless, for illustrative and demonstrative consistency and simplicity, as well as for their popularity in the literature, throughout this thesis, triangular MFs are employed for all FRI methods.

As shown in Fig. 2.1, a normal and convex triangular fuzzy set A is illustrated with its three ascending-ordered CPs, i.e., (a_1, a_2, a_3), where the first and third CP stand for the two extreme points of the support with a membership value of 0 and the middle one stands for the normal point of the fuzzy set with a membership of 1. For a fuzzy rule base consisting of rules in the form as per Eqn. (2.1), the triangular fuzzy values A_j^i, A_j^*, B^i, and the consequent B^* to be computed by an FRI process $(i = 1, 2, \ldots, N, j = 1, 2, \ldots, m)$ are therefore, represented by their corresponding CPs: $(a_{j1}^i, a_{j2}^i, a_{j3}^i)$, $(a_{j1}^*, a_{j2}^*, a_{j3}^*)$, (b_1^i, b_2^i, b_3^i), and (b_1^*, b_2^*, b_3^*), respectively.

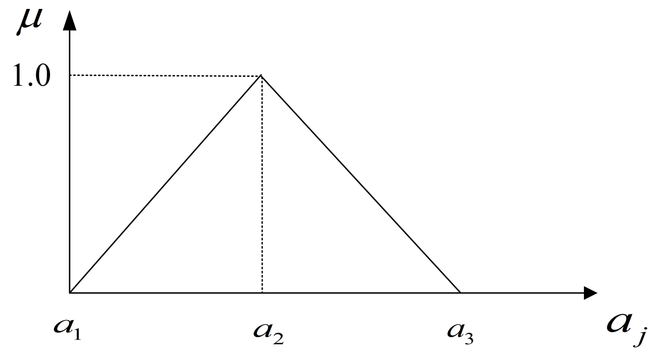


Figure 2.1: Normal and convex triangular membership function.

2.2 Fuzzy Rule Interpolation Techniques

Fuzzy rule bases are the essential component of any approximate reasoning model. Their properties determine specifically what techniques to use in order to accomplish the required inference. Conventional fuzzy inference mechanism, represented by Zadeh and Mamdani's compositional rule of inference (CRI), has been successfully applied to many problems. However, it is significantly restricted in situations where dense rule bases are not available. That is, the problem domain is not completely covered by the given rules where certain observation do not overlap with any rules (fully or partially to a satisfactory degree).

In many circumstances, a dense rule base cannot be realistically obtained, but only an incomplete rule base instead. A number of reasons may lead to such incomplete rule bases, for example, the most common reasons include [Baranyi et al., 1999, Tikk and Baranyi, 2000]:

- To utilise incomplete knowledge about the modelled problem, regardless of the means for the construction of the rule base, be it from human expertise or machine learning techniques; and
- To reduce the number of rules in a rule base and hence, the complexity of the resultant fuzzy system.

As briefly outlined in Chapter 1, FRI techniques enable fuzzy interpolative reasoning to be performed with sparse knowledge. This section categorises and details the representatives of classical FRI methods, with typical pros and cons of different approaches discussed.

2.2.1 Categorisation of FRI Approaches

In the literature, various FRI approaches have been proposed following the seminal work of [Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b], to perform fuzzy interpolative reasoning. In general, the existing methodologies can be grouped into two categories:

1. α-cut / Non-transformation based FRI, see Table 2.1 for a summary with Table 2.2 listing further developments belonging to this category.
2. Intermediate rule / Transformation based FRI, see Table 2.3 for a summary with Table 2.4 listing a particular family of scale and move transformation based FRI (denoted as T-FRI hereafter) which are the most popular in the recent literature.

This categorisation is made depending upon whether processes to construct and then to utilise a so-called intermediate fuzzy rule are involved in order to derive an interpolated result.

The α-cut based FRI approaches, also known as non-transformation based methods, directly interpolate the results based on the computation of each α-cut level given at least two fuzzy rules adjacent to an unmatched observation. Considerable work has been reported on this type of approach at the early stage of the investigation of fuzzy interpolative reasoning. In particular, the very first proposed, termed the KH method after the name of its inventors [Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b], is the most typical α-cut based algorithm for FRI. As indicated earlier, Table 2.1 also summarises several other alternative α-cut based methods from different perspectives.

Table 2.1: α-cut (Non-transformation) based FRI Methods

Methods	Characteristics
[Huang et al., 2004, Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b, Ughetto et al., 2000]	FRI with only two fuzzy rules
[Chang et al., 2008, Chen and Chen, 2016, Chen et al., 2013a, Chen et al., 2015, Chen and Lee, 2011, Cheng et al., 2015, Cheng et al., 2016, Kovács, 2006, Yang and Shen, 2013]	FRI with multiple fuzzy rules
[Chen and Chen, 2016, Chen et al., 2013a, Cheng et al., 2015]	FRI with fuzzy rules weighted
[Chen and Lee, 2011]	FRI with interval type-2 fuzzy sets [Mendel et al., 2006]
[Chen et al., 2015]	FRI with rough-fuzzy sets
[Chen and Adam, 2017, Cheng et al., 2016]	FRI with adaptivity

Table 2.2: Family of KH FRI

Methods	Characteristics
[Koczy and Hirota, 1997, Kóczy et al., 2000, Koczy et al., 2000, Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b, Kóczy and Hirota, 1993c, Kóczy and Hirota, 1991]	Foundational linear KH FRI based on computation of α-cut levels
[Vass et al., 1992]	Extended KH FRI with reduction of invalid conclusions
[Baranyi et al., 1999, Tikk, 1999, Tikk and Baranyi, 2000, Yam et al., 1999]	Modified α-cut based method based on coordinate modification
[Tikk et al., 1997, Tikk et al., 1999, Tikk et al., 2002]	Stabilised (general) KH interpolation
[Wong et al., 2000, Wong et al., 2005]	Modified α-cut based multidimensional scheme

For the group of transformation-based approaches, they work by first computing an intermediate rule. The required consequent to an unknown observation is obtained through a two-step procedure by manipulating selected neighbouring rules to the observation. An intermediate rule is artificially constructed such that its antecedent is as “close” (given a certain distance metric, often the Euclidean one) to the observation as possible. An intermediate consequent is computed from the constructed rule antecedent. Observing that there may still exist a difference between the antecedent of the intermediate rule and the observation, the second step works based on the principle of analogical reasoning mechanism [Bouchon-Meunier and Valverde, 1999, Turksen and Zhong, 1988]. It derives the conclusion by transforming the intermediate consequent in terms of the similarity measured between the antecedent of the intermediate rule and the observation, in an analogical manner as transforming the intermediate rule antecedent to the given observation. Note that the foundational T-FRI methodology, as one of the outstanding intermediate rule based FRI methods, has first introduced in [Huang and Shen, 2006, Huang and Shen, 2008], which also forms the basis for the work presented in this thesis. Many follow-on developments and modifications to this seminal approach have been proposed over the last two decades.

Table 2.3: Intermediate Rule (Transformation) based FRI Methods

Methods	Characteristics
[Hsiao et al., 1998]	Exploiting slopes of fuzzy sets to obtain valid conclusions
[Wu et al., 1996]	Using similarity transfers to guarantee valid interpolation
[Baranyi et al., 1995, Baranyi et al., 1996a, Baranyi and Kóczy, 1996a, Baranyi et al., 2004, Baranyi et al., 1996b, Baranyi et al., 1998, Baranyi and Kóczy, 1996b]	Adopting generalised concept for interpolation and extrapolation
[Kawaguchi and Miyakoshi, 2000a, Kawaguchi and Miyakoshi, 2000b, Kawaguchi and Miyakoshi, 2001, Kawaguchi et al., 1997]	Performing B-spline based interpolation
[Chen et al., 2016, Chen and Shen, 2017, Huang and Shen, 2006, Huang and Shen, 2008, Jin et al., 2014, Li et al., 2019a, Li et al., 2018c, Naik et al., 2017b, Shen and Yang, 2011, Yang et al., 2017, Yang and Shen, 2011]	Running FRI with scale and move transformation (T-FRI)

Apart from the two major groups of FRI methods to conduct fuzzy interpolative reasoning as outlined above, there are alternative FRI techniques, as summarised in Table 2.5. This shows the diversity of this interesting research area.

Table 2.4: Family of Scale and Move Transformation based FRI (T-FRI)

Methods	Characteristics
[Huang and Shen, 2006]	Foundational T-FRI working with two given neighbouring rules involving multiple antecedent variables in various fuzzy membership functions (e.g., complex polygon, Gaussian or bell-shaped)
[Huang and Shen, 2008]	Extended T-FRI facilitating both interpolation and extrapolation involving multiple fuzzy rules, with each rule consisting of multiple antecedents
[Jin et al., 2014, Jin et al., 2019]	Backward T-FRI allowing missing antecedent values directly related to the consequent to be interpolated from known antecedents and consequent, supporting backward interpolation and extrapolation involving multiple multi-antecedent fuzzy rules
[Yang et al., 2017,Yang and Shen, 2011]	Adaptive T-FRI being capable of restoring system consistency once contradictory results reached during interpolation
[Chen et al., 2016, Chen and Shen, 2012]	Rough-fuzzy T-FRI allowing representation, handling and utilisation of different levels of uncertainty in knowledge
[Chen and Shen, 2017]	Extended T-FRI with interval type-2 fuzzy sets
[Naik et al., 2014,Naik et al., 2017b]	Dynamic T-FRI facilitating selection, combination, and generalisation of informative, frequently used interpolated rules for enriching existing rule base while performing interpolation

The efficacy of the inference mechanism introduced by an FRI method may be reflected or revealed through their utilisation in resolving real-world application problems. As with many classic fuzzy reasoning tools, FRI has particularly reinforced the power of systems control, including successful examples for: simulation of automated guided vehicles [Kovács and Kóczy, 1999], surveillance navigation control of mobile robots [Vincze and Kovács, 2008], and general behaviour-based control [Kovács and Kóczy, 2004]. The work on dynamic FRI [Naik et al., 2017b] has offered significant opportunities for facilitating selection, combination and generalisation of informative, frequently used interpolated rules for enriching existing rule bases while performing interpolation. It provides promising solutions to cyber-security problems, including: network security analysis, intelligent intrusion detection [Naik et al., 2017b] and firewall reinforcement (especially for Microsoft Windows Firewall) [Naik et al., 2017a]. FRI also finds impressive results in performing practical pattern recognition

Table 2.5: Alternative FRI Techniques

Methods	Characteristics
[Kovács and Kóczy, 1997, Kovacs and Koczy, 1997, Kovács and Kóczy, 1997]	Interpolation based on approximation of vague environment of fuzzy rules with application to automatic guided vehicle systems
[Bouchon-Meunier et al., 1999, Bouchon-Meunier et al., 2001, Bouchon-Meunier et al., 2000]	Interpolative method based on graduality
[Jenei, 2001, Jenei et al., 2002]	Axiomatic approach for interpolation and extrapolation of fuzzy quantities
[Yam et al., 2000b, Yam and Kóczy, 2000, Yam and Kóczy, 1998, Yam and Kóczy, 2001, Yam et al., 2000a]	Cartesian based interpolation with each fuzzy set mapped onto a point in high dimensional Cartesian space

tasks, examples include: classic prediction problem [Chen and Chen, 2016] using weighted FRI techniques; computer vision and image super resolution [Yang et al., 2019]; and disease diagnosis in general and colorectal polyp detection [Nagy et al., 2018] in particular. Further applications of FRI are found in function approximation [Wong and Gedeon, 2000, Berecz, 2009] and student academic performance evaluation [Johanyák, 2010].

For the purpose of demonstrating the basic ideas of typical FRI methods, and more importantly, for the following development of any weighted version that improves on the original FRI (where no individual weights of rule antecedent variables are involved), several commonly used FRI approaches from each of the two main categories are reviewed below. As indicated previously, the triangular fuzzy membership functions, as defined in Section 2.1 are employed throughout, unless otherwise stated, both for consistency in demonstration of the ideas and for efficiency in computation.

2.2.2 Representative α-Cut based FRI

The α-cut based interpolation is essentially a fuzzy extension of the classical linear interpolation of given points that are linked with fuzzy rules. The interpolated result is generated through the computation and then, the aggregation of linear interpolation at each α-cut level. Theoretically, in the case of arbitrary shaped convex normalised fuzzy sets an infinite number of α-levels should be taken into consideration for an approximate conclusion. In practice however, to achieve an acceptable computational requirement, most α-cut based methods only take a finite number of α-levels (usually two, three or four) into account, with the resulting points being connected piecewise linearly to yield an approximation of the consequent.

2.2.2.1 KH: Foundational Linear FRI

This section first formulates the basic idea of the most famous α-cut based FRI, named KH linear FRI (after its inventors [Kóczy and Hirota, 1993a, Kóczy and Hirota, 1993b]), in a general formation, followed by its practical implementation by the use of triangular membership functions in a multidimensional situation.

A. Core Principle

The KH rule interpolation offers an initial proposal for fuzzy interpolative reasoning through manipulating α-cut distances. When a given observation fails to match any rule in the sparse rule base for firing, an interpolated consequent is constructed by performing a linear aggregation of the rule consequents of a number (usually two) of selected neighbouring rules closest to the observation. Such failure for activating rule(s) may be generally due to no matching, or in certain FRI-based systems, due to too low level a partial matching. The above aggregation operation complies with the general principle of similarity-based analogical reasoning, such that

The closer a rule's antecedent A^i (which is a logical aggregation of individual attribute values A^i_j) to the observation o^*, the closer the rule's consequent B^i to the outcome B^* that corresponds to o^*.

The similarity measure employed is specified by the use of fuzzy distances defined between a rule antecedent and the observation. That is, the smaller distance between A^i and o^* is, the more similar they are, with the corresponding B^i deemed to potentially make more contribution than otherwise to the consequent being sought.

Suppose that there are two rules r^i and r^j in the rule base R, which are formulated as shown in Eqn. (2.1). Given an observation o^* (again, as per Eqn. (2.1)), the notion of linear rule interpolation can be written as:

$$\frac{\tilde{d}(A^*,A^i)}{\tilde{d}(A^*,A^j)} = \frac{\tilde{d}(B^*,B^i)}{\tilde{d}(B^*,B^j)} \tag{2.2}$$

where

$$\tilde{d}(A^*,A^i) = \sqrt{\tilde{d}_{i1}^2 + \tilde{d}_{i2}^2 + \cdots + \tilde{d}_{im}^2} \qquad \tilde{d}_{it} = \tilde{d}(A_t^*,A_t^i), t = 1,2,\cdots,m \tag{2.3}$$

and $\tilde{d}$ denotes the fuzzy distance between the two membership functions.

Fuzzy distance between two fuzzy sets is interpreted as a pair of *lower and upper fuzzy distances* between their α-cut sets, with respect to the *Resolution Principle* [Kóczy and Hirota, 1993a]. For a particular $\alpha \in [0,1]$, the lower fuzzy distance $\tilde{d}_L(A,B)$ and upper fuzzy distance $\tilde{d}_U(A,B)$ are denoted as:

$$\tilde{d}_L(A,B) = D(inf(A_\alpha), inf(B_\alpha)) \qquad \tilde{d}_U(A,B) = D(sup(A_\alpha), sup(B_\alpha)) \tag{2.4}$$

where D denotes the Minkowski distance, and $inf(\cdot)$ and $sup(\cdot)$ are the infimum and supremum of the α-cut concerned, respectively. Hence, the formula of linear rule interpolation (i.e., Eqn. (2.2)) can be rewritten as:

$$\begin{aligned} \frac{\tilde{d}_L(A_\alpha^*,A_\alpha^i)}{\tilde{d}_L(A_\alpha^*,A_\alpha^j)} &= \frac{\tilde{d}_L(B_\alpha^*,B_\alpha^i)}{\tilde{d}_L(B_\alpha^*,B_\alpha^j)} \\ \frac{\tilde{d}_U(A_\alpha^*,A_\alpha^i)}{\tilde{d}_U(A_\alpha^*,A_\alpha^j)} &= \frac{\tilde{d}_U(B_\alpha^*,B_\alpha^i)}{\tilde{d}_U(B_\alpha^*,B_\alpha^j)} \end{aligned} \tag{2.5}$$

This leads to the solution for $min\{B_\alpha^*\}$ and $max\{B_\alpha^*\}$ being:

$$\begin{aligned} min\{B_\alpha^*\} &= \frac{w_{\alpha L}^i min\{B_\alpha^i\} + w_{\alpha L}^j min\{B_\alpha^j\}}{w_{\alpha L}^i + w_{\alpha L}^j} \\ max\{B_\alpha^*\} &= \frac{w_{\alpha U}^i max\{B_\alpha^i\} + w_{\alpha U}^j max\{B_\alpha^j\}}{w_{\alpha U}^i + w_{\alpha U}^j} \end{aligned} \tag{2.6}$$

where

$$
\begin{aligned}
w^i_{\alpha L} &= \frac{1}{\tilde{d}_L(A^*_\alpha, A^i_\alpha)} & w^j_{\alpha L} &= \frac{1}{\tilde{d}_L(A^*_\alpha, A^j_\alpha)} \\
w^i_{\alpha U} &= \frac{1}{\tilde{d}_U(A^*_\alpha, A^i_\alpha)} & w^j_{\alpha U} &= \frac{1}{\tilde{d}_U(A^*_\alpha, A^j_\alpha)}
\end{aligned}
\tag{2.7}
$$

The α-cut of conclusion is then given by

$$
B^*_\alpha = [min\{B^*_\alpha\}, max\{B^*_\alpha\}] \tag{2.8}
$$

and the interpolated conclusion can therefore, be obtained by the use of *Resolution Principle* such that

$$
B^* = \bigcup_{\alpha \in [0,1]} B^*_\alpha \tag{2.9}
$$

B. Multidimensional Implementation

The foundational KH FRI works effectively and efficiently for simple linear problems. It has been subsequently developed to address sparse rule interpolation in more complex situations, for instance involving multiple rules with multiple antecedent variables [Tikk et al., 2002, Wong et al., 2005]. Thanks to the piecewise linear property presumed by KH interpolation, given triangular membership functions, the interpolated outcome $B^* = (b^*_1, b^*_2, b^*_3)$ can be determined with its two α-cut sets (when α is 0 or 1), resulting in the three characteristic points taking the values of

$$
b^*_t = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{\sum_{j=1}^{m} (a^i_{jt} - a^*_{jt})^2}} b^i_t}{\sum_{i=1}^{n} \frac{1}{\sqrt{\sum_{j=1}^{m} (a^i_{jt} - a^*_{jt})^2}}} \tag{2.10}
$$

where n is the number of the neighbouring rules used for interpolation, m is the number of attributes in the rule, and $t = 1, 2, 3$. Such computation for the interpolated fuzzy set B^* reflects exactly the general situation as expressed by Eqn. (2.6), where

$$
\begin{aligned}
b_1^* &= min\{B_0^*\} \\
b_3^* &= max\{B_0^*\}, \qquad \alpha = 0 \\
b_2^* &= min\{B_1^*\} = max\{B_1^*\}, \qquad \alpha = 1
\end{aligned} \tag{2.11}
$$

2.2.2.2 CCL Rule Interpolation

As one of the most popularly used α-cut based FRI methods, the CCL rule interpolation (named after its inventors [Chang et al., 2008]) offers an alternative means for fuzzy interpolative reasoning that exploits the areas of the fuzzy sets involved in the rules and the (unmatched) observation. The idea is to preserve the logically consistent properties with respect to the *ratio of fuzziness* (RF), which is determined by the areas of the fuzzy sets concerned. That is, it pursues consistency of RF between the (to be) interpolated consequent over the observation and the consequent value over the antecedent value of each rule used for interpolation. More specifically, the RF between two fuzzy values A and B is defined by

$$RF(A,B) = \frac{S(A)}{S(B)} \tag{2.12}$$

where $S(A), S(B)$ denote the area of the fuzzy set of A and that of B, respectively.

The CCL FRI method presents a flexible interpolative reasoning framework, allowing the use of different types of membership function (MF), including various polygonal typed and Gaussian shaped MFs. It can also handle general cases that involve multiple antecedent variables involved in multiple fuzzy rules. For simplification and consistency throughout, the core computations are summarised below in relation to the use of triangular fuzzy membership functions.

First, the normal point b_2^* of the (to be) interpolated consequent B^* is defined by linear interpolation, such that

$$b_2^* = \sum_{i=1}^{n} W_i b_2^i \tag{2.13}$$

$$S_K(B^*) = \begin{cases} \left(\sum_{j=1}^{m} S_K(A_j^*)\right) \times \left(\sum_{\substack{i=1, \\ \exists j S_K(A_j^i)>0}}^{n} W_i \times \frac{S_K(B^i)}{\sum_{j=1}^{m} S_K(A_j^i)} \right), \text{ if } \exists ij S_K(A_j^i) > 0 \\ \frac{\sum_{j=1}^{m} S_K(A_j^*)}{m}, \text{ if } \forall ij S_K(A_j^i) = 0 \end{cases} \quad (2.14)$$

in which n is the number of selected rules for interpolation, and W_i is the aggregated rule weight, which is calculated by

$$W_i = \frac{\sum_{j=1}^{m} w_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}}, \qquad w_{ij} = 1 - \left| \frac{a_{j2}^i - a_{j2}^*}{max_{a_{j2}} - min_{a_{j2}}} \right| \quad (2.15)$$

where $max_{a_{j2}}$ and $min_{a_{j2}}$ are used for normalisation, denoting the maximal and minimal value within $\{a_{j2}^i | i = 1, 2, \ldots, n\}$.

Given the three characteristic points created from the two α-cut sets (when $\alpha = 0, 1$), a triangular fuzzy set is divided into two smaller sub-triangles, as shown in Fig. 2.2 (more triangular or even trapezoidal shaped sub-polygons may be generated for more complex polygonal fuzzy sets with many characteristic points, but the same idea is followed as herein). From this, the left triangular area $S_L(B^*)$ (i.e., the part of the geometrical area of a triangular fuzzy set on the left hand side of the normal point) and the right triangular area $S_R(B^*)$, of the fuzzy set B^* are calculated by Eqn. (2.14), where for the subscript S_K, $K \in \{L, R\}$. This equation exactly reveals the basic idea of the CCL rule interpolation, where the RF from the observation viewpoint is constructed by the weighted aggregation of the RF of the involved rules, thereby leading to the derivation of the area of the interpolated fuzzy set.

Finally, the left and right extreme points of the support for the interpolated result B^* are derived from the resulting triangular areas as follows:

$$b_1^* = b_2^* - 2S_L(B^*), \qquad b_3^* = b_2^* + 2S_R(B^*) \quad (2.16)$$

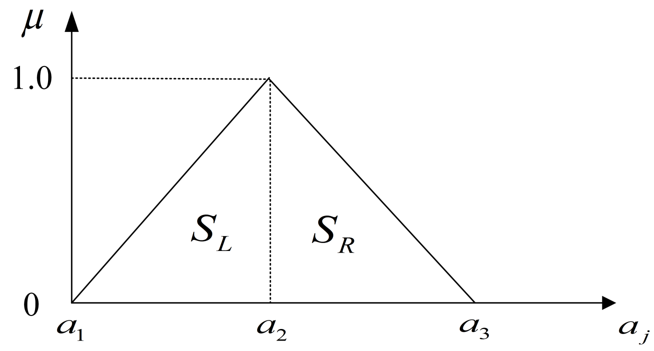


Figure 2.2: Left area S_L and right area S_R of triangular fuzzy set.

2.2.3 Representative Intermediate Rule based FRI

This section reviews the underlying interpolation mechanism of the intermediate rule (or transformation) based FRI. In particular, more detailed description is given to the scale and move transformation-based FRI approach which has the closest relation to the research described in this thesis.

2.2.3.1 Representative Value of Fuzzy Set

Prior to going through the details of intermediate rule based FRI techniques, a very important concept needs to be introduced, which is adopted within this type of interpolation algorithm. This is the *Representative Value (Rep)* of a fuzzy set. There are actually many variations in the literature (e.g., [Baranyi et al., 2004, Chen and Chang, 2011b, Chen et al., 2009, Huang and Shen, 2008]), assigned with different names, such as representative value [Huang and Shen, 2008], reference point [Baranyi et al., 2004], and characteristic value [Chen and Ko, 2008, Chen et al., 2009]. Nonetheless, they imply similar interpretations as described below with the term of representative value.

The representative value of a fuzzy set is a single value assigned to help capture important information contained by the set in a simplified way, such as the "most typical" overall location of the fuzzy set in its domain and also, its geometric shape. In certain situations, the *Rep* value may be defined by the defuzzified value of the fuzzy set if that is preferred since there is no unified definition. What is important is within a particular FRI method, all *Rep* values are computed in the same way.

More formally as with the most popular approach in the literature, given an arbitrary polygonal fuzzy set $A = (a_1, a_2, \ldots, a_n)$ where $a_i, i = 1, 2, \ldots, n$ denote the characteristic points of the polygonal, its representative value $Rep(A)$ is defined by:

$$Rep(A) = \sum_{i=1}^{n} w_i a_i \tag{2.17}$$

where w_i is the weight assigned to the characteristic point a_i per i. In particular, the simplest case, which is named the average *Rep*, is one so computed where all points take the same weight value, i.e., $w_i = 1/n$.

For computational simplicity, many fuzzy rule-based systems (including the present work) have adopted triangular membership functions to define fuzzy sets while representing attribute values. For such a fuzzy set $A = (a_1, a_2, a_3)$ given in Fig. 2.1 of Section 2.1, $Rep(A)$ is simply defined as follows (though its centre of gravity may also be used as an alternative if preferred):

$$Rep(A) = \frac{a_1 + a_2 + a_3}{3} \tag{2.18}$$

The definition of representative values for more complex membership functions can be found in [Huang and Shen, 2008].

Apart from its geometrical meaning, the *Rep* value also simplifies the definition of the distance between fuzzy sets, to measure the degree of "closeness". A simple case of the distance between two fuzzy sets A and B can be defined by

$$d(A, B) = \left| Rep(A) - Rep(B) \right| \tag{2.19}$$

which is a crisp distance in contrast with α-cut distance based methods [Baranyi et al., 2004]. The distance definition employed in a given FRI approach will be specified in each method later.

2.2.3.2 Scale and Move Transformation based FRI (T-FRI)

The scale and move transformation based FRI (T-FRI) is one of the most general and advanced intermediate rule based FRI mechanisms. One of the key aims of this development has been to eliminate an important practical issue that earlier work of FRI had in that the interpolated outcomes were not guaranteed to be convex and in certain cases, not even a fuzzy set. The presentation of the fundamental idea of T-FRI is reported in [Huang and Shen, 2006, Huang and Shen, 2008], which can handle both interpolation and extrapolation of multiple multi-antecedent rules with complex polygon shaped, Gaussian and bell-shaped fuzzy membership functions. The following outlines the key computational steps of T-FRI working with multiple fuzzy rules where in general, multiple rule antecedents are involved in each rule.

Given a sparse rule base R and an observation o^*, in the form of Eqn. (2.1), T-FRI works by running a computational process as highlighted in Fig. 2.3, involving four core procedures as summarised below.

A. Selection of Closest Rules

This procedure is required as the basis upon which to perform FRI, when o^* does not match any of the rules in the rule base. Intuitively, it searches for a certain number of rules that are closest to the observation. The distance between an observation o^* and a rule r^q, or the distance between any two rules $r^p, r^q \in R$, is determined by computing the aggregated distances over all the corresponding fuzzy values of the shared attributes between them:

$$d(v, r^q) = \frac{1}{\sqrt{m}} \sqrt{\sum_{j=1}^{m} d(A_j^v, A_j^q)^2} \tag{2.20}$$

where v is o^* or r^p (so A_j^v is A_j^* or A_j^p), depending on whether the distance is between an observation and a rule or between two rules. So, the n closest rules to o^* are those rules leading to the n smallest values of this distance measurement.

In the above definition,

$$d(A_j^v, A_j^q) = \frac{\left| Rep(A_j^v) - Rep(A_j^q) \right|}{max_{A_j} - min_{A_j}} \tag{2.21}$$

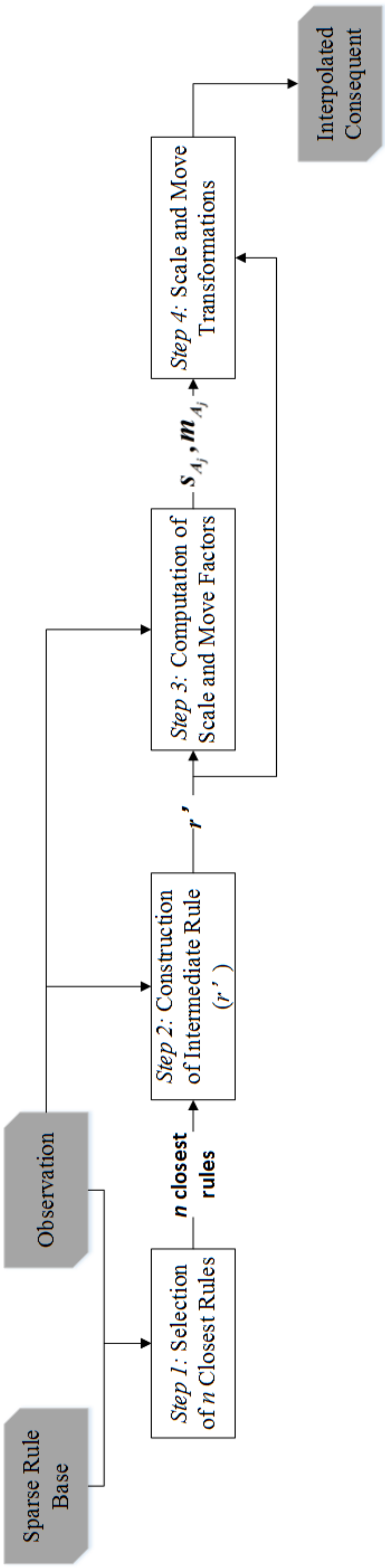


Figure 2.3: Framework of transformation-based FRI.

This is implemented by the use of the *Rep* values of the corresponding fuzzy sets (defined in Section 2.2.3.1), representing the normalised result of the otherwise absolute distance, where max_{A_j} and min_{A_j} denote the maximal and minimal value of the attribute a_j, respectively. This normalisation is to ensure that all distance measures are compatible with each other over different attribute domains.

B. Construction of Intermediate Fuzzy Rule

From the preceding procedure, n closest rules to a given observation can be chosen which have the minimal distances amongst all the rules with respect to the observation. From this, an intermediate fuzzy rule r' is constructed, forming the start point of the transformation process in T-FRI. In most applications of T-FRI, n is taken to be 2 purely for computational efficiency, but often at the expense of interpolative accuracy (if all rule antecedents are regarded as of having equal significance).

The construction procedure computes the antecedent fuzzy sets $A'_j, j = 1, \ldots, m$ and the corresponding consequent fuzzy set B' of the intermediate rule:

r' : if a_1 is A'_1 and a_2 is A'_2 and $\cdots$ and a_m is A'_m, then z is B'

which is a weighted aggregation of the n closest rules. Let $w^i_j, i \in \{1, \ldots, n\}$, denote the weight to which the jth antecedent of the ith fuzzy rule contributes to the construction of the jth antecedent A'_j of the intermediate fuzzy rule:

$$w^i_j = \frac{1}{1 + d(A^i_j, A^*_j)} \tag{2.22}$$

where $d(A^i_j, A^*_j)$ is calculated as per (2.21). Then,

$$A'_j = A''_j + \delta_{A_j}(max_{A_j} - min_{A_j}) \tag{2.23}$$

with

$$A''_j = \sum_{i=1,\ldots,n} \hat{w}^i_j A^i_j \tag{2.24}$$

where $\hat{w}^i_j$ denotes the normalised weight and δ_{A_j} is a constant (termed the shift factor of A_j), defined respectively by

$$\hat{w}^i_j = \frac{w^i_j}{\sum_{t=1,\ldots,n} w^t_j} \tag{2.25}$$

$$\delta_{A_j} = \frac{|Rep(A^*_j) - Rep(A''_j)|}{max_{A_j} - min_{A_j}} \tag{2.26}$$

The consequent value of the intermediate rule is constructed in the same manner as above, that is

$$B' = B'' + \delta_z(max_z - min_z) \tag{2.27}$$

where max_z and min_z are the maximal and minimal values of the consequent attribute, B'' is the weighted aggregation of the consequent values of the n closest rules $B^i, i = 1, \ldots, n$:

$$B'' = \sum_{i=1,\ldots,n} \hat{w}^i_z B^i \tag{2.28}$$

with $\hat{w}^i_z$ being the mean of the normalised weights associated with the antecedents $\hat{w}^i_j$ in each rule:

$$\hat{w}^i_z = \frac{1}{m} \sum_{j=1}^{m} \hat{w}^i_j \tag{2.29}$$

and the shift factor δ_z of the consequent is the mean of $\delta_{A_j}, j = 1, \ldots, m$

$$\delta_z = \frac{1}{m} \sum_{j=1}^{m} \delta_{A_j} \tag{2.30}$$

C. Computation of Scale and Move Factors

The goal of a transformation process T in T-FRI is to scale and move an intermediate fuzzy set A'_j, such that the transformed shape and representative value coincide with those of the observed value A^*_j. This process is implemented in two stages:

1. scale operation from A'_j to $\hat{A}'_j$ (denoting the scaled intermediate fuzzy set); and
2. move operation from $\hat{A}'_j$ to A^*_j.

For this purpose, the required scale rate s_{A_j} and move ratio m_{A_j} are determined in this step. It computes and records all such scale rates and move ratios for use in the subsequent, and final, procedure to obtain the required consequent value. Unfortunately, it is difficult to have a generic, closed form representation of these transformation factors as they are dependent upon the fuzzy membership functions used.

For this work, triangular fuzzy sets are used throughout. Given such a fuzzy set $A'_j = (a'_{j1}, a'_{j2}, a'_{j3})$, the scale rate s_{A_j} is:

$$s_{A_j} = \frac{a^*_{j3} - a^*_{j1}}{a'_{j3} - a'_{j1}} \tag{2.31}$$

which essentially expands or contracts the support length of $A'_j : a'_{j3} - a'_{j1}$ so that it becomes the same as that of A^*_j. The scaled intermediate fuzzy set $\hat{A}'_j$, which has the same representative value as A'_j, is then obtained such that

$$\begin{aligned}
\hat{a}'_{j1} &= \frac{(1+2s_{A_j})a'_{j1} + (1-s_{A_j})a'_{j2} + (1-s_{A_j})a'_{j3}}{3} \\
\hat{a}'_{j2} &= \frac{(1-s_{A_j})a'_{j1} + (1+2s_{A_j})a'_{j2} + (1-s_{A_j})a'_{j3}}{3} \\
\hat{a}'_{j3} &= \frac{(1-s_{A_j})a'_{j1} + (1-s_{A_j})a'_{j2} + (1+2s_{A_j})a'_{j3}}{3}
\end{aligned} \tag{2.32}$$

Similarly, while dealing with triangular fuzzy sets, the move operation shifts the position of $\hat{A}'_j$ to becoming the same as that of A^*_j, while maintaining its representative value $Rep(\hat{A}'_j)$. This is achieved using the move ratio m_{A_j}:

$$m_{A_j} = \begin{cases} \frac{3(a^*_{j1} - \hat{a}'_{j1})}{\hat{a}'_{j2} - \hat{a}'_{j1}}, \text{ if } a^*_{j1} \geq \hat{a}'_{j1} \\ \frac{3(a^*_{j1} - \hat{a}'_{j1})}{\hat{a}'_{j3} - \hat{a}'_{j2}}, \text{ otherwise} \end{cases} \tag{2.33}$$

D. Scale and Move Transformation

After calculating the necessary scale and move factors (i.e., s_{A_j} and $m_{A_j}, j = 1, \ldots, m$), this procedure completes the T-FRI process, deriving the required consequent value of B^*. This follows the intuition of similar observations leading to similar consequents, a heuristic fundamental to analogical approximate reasoning. For this, the transformation factors on the antecedent attributes are aggregated. In all conventional T-FRI methods, this is implemented by averaging them:

$$s_z = \frac{1}{m}\sum_{j=1}^{m} s_{A_j} \qquad m_z = \frac{1}{m}\sum_{j=1}^{m} m_{A_j} \tag{2.34}$$

This entails the computation of scaled $\hat{B}' = (\hat{b}'_1, \hat{b}'_2, \hat{b}'_3)$:

$$\begin{aligned} \hat{b}'_1 &= \frac{(1+2s_z)b'_1 + (1-s_z)b'_2 + (1-s_z)b'_3}{3} \\ \hat{b}'_2 &= \frac{(1-s_z)b'_1 + (1+2s_z)b'_2 + (1-s_z)b'_3}{3} \\ \hat{b}'_3 &= \frac{(1-s_z)b'_1 + (1-s_z)b'_2 + (1+2s_z)b'_3}{3} \end{aligned} \tag{2.35}$$

where $B' = (b'_1, b'_2, b'_3)$ is the fuzzy value of the intermediate consequent previously computed. From this, again, by analogy to the transformation required for the antecedent to match the observation, move transformation is applied, resulting in the final, required interpolated consequent $B^* = (b^*_1, b^*_2, b^*_3)$:

$$\begin{aligned} b^*_1 &= \hat{b}'_1 + m_z\gamma \\ b^*_2 &= \hat{b}'_2 - 2m_z\gamma \qquad \gamma = \begin{cases} \frac{\hat{b}'_2 - \hat{b}'_1}{3}, \text{ if } m_z \geq 0 \\ \frac{\hat{b}'_3 - \hat{b}'_2}{3}, \text{ otherwise} \end{cases} \\ b^*_3 &= \hat{b}'_3 + m_z\gamma \end{aligned} \tag{2.36}$$

For illustration, Fig. 2.4 outlines the scale and move transformation process (i.e., Steps C and D of a typical T-FRI method), where the scale and move factors of each rule antecedent are shown to be calculated in the upper box and the interpolated result is obtained by the corresponding transformations shown underneath. For conciseness, such a process can be collectively represented by: $B^* = T(B', s_z, m_z)$, emphasising on the significance of both scale and move transformations.

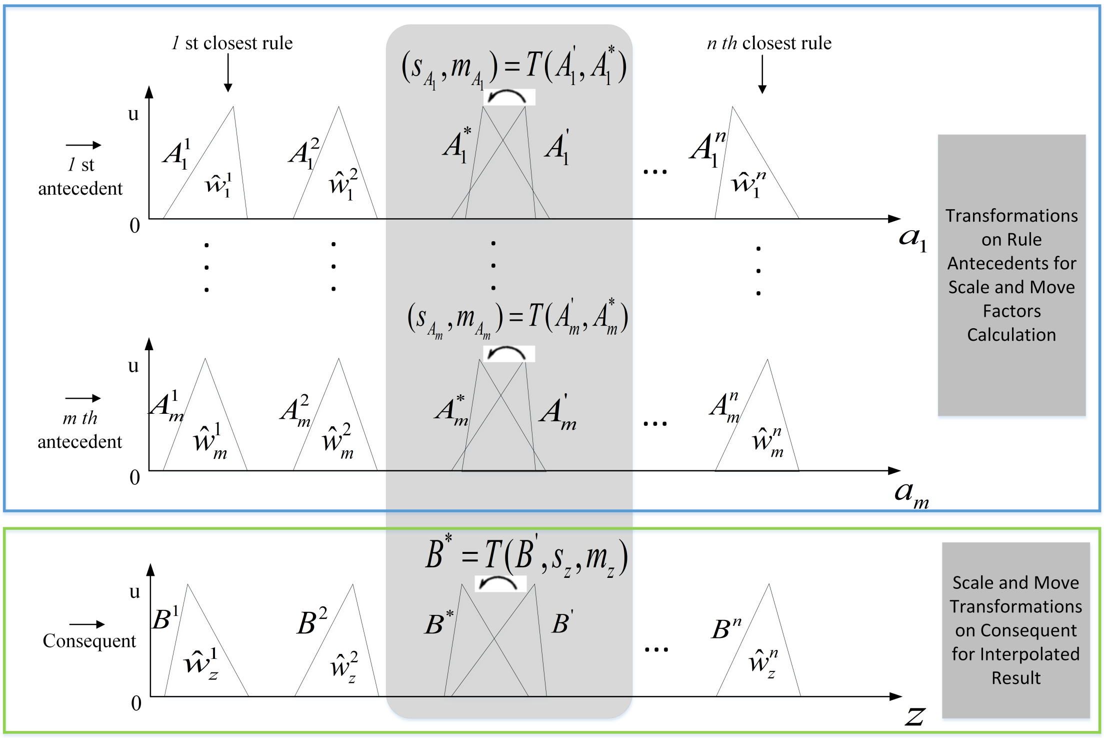


Figure 2.4: Fuzzy rule interpolation via scale and move transformations.

2.2.3.3 Representative Modifications to Scale and Move Transformation-based FRI

Following the generic and seminal ideas of the above-reviewed T-FRI approach, there have been a large family of works that have been proposed to further improve it, of which an overview is previously shown in Table 2.4. This section provides an outline of representative methods within this family.

- Adaptive T-FRI [Yang and Shen, 2011, Yang et al., 2017]

This work is motivated by an observation that there may exist inconsistency in interpolated results after a sequence of T-FRI operations. The potential reasons have been analysed to include detective interpolated rules and inaccurate interpolative transformations. The adaptive fuzzy interpolation enhances the original T-FRI with the ability for identification and correction of defective rules in interpolative transformations, facilitating the removal of certain inconsistencies. This is accomplished through two sub-systems: 1) a diagnostic sub-system that is constructed by the use of the general diagnostic engine, where the inconsistent interpolated results are recorded in an ATMS (assumption-based truth maintenance system) [De Kleer, 1986]; and 2) a corrective sub-system that is derived from a fuzzy extension to the traditional numerical interpolation theory and its application in approximation computation. However, this work is focussed on the implementation of adaptive T-FRI that involves just two multiple-antecedent rules. Besides, further investigation is required to reveal whether it can handle situations where extrapolation is necessary (since the original T-FRI is able to deal with extrapolation in the same manner as with interpolation).

- Backward T-FRI [Jin et al., 2014, Jin et al., 2019]

Conventional FRI generally executes in a "forward" manner, where the consequent is required to be interpolated given the rule base and all antecedent attributes of an observation available. Nevertheless, situations may arise when certain crucial antecedents are absent from the given observation, which may also be involved in the subsequent interpolation process, thereby leading to the failure of the derivation of the final interpolated conclusion. This important issue is addressed by a modification of T-FRI, termed backward T-FRI, which provides a series of solutions for handling both single missing antecedent value and multiple missing antecedents problems. The single missing antecedent issue is resolved by implementing a four-step computation procedures (mirroring what is presented in Section 2.2.3.2) of the original T-FRI, resulting in the reverse calculation of the relative parameters corresponding to the unknown antecedent value. The general backward T-FRI with multiple missing antecedent values is addressed by two procedures: 1) a direct extension of the method for the single missing value case, by estimating parameter combinations that would lead to the closest resemblance of the original (missing) values; and 2) an approach for the removal of possible missing antecedent values through a process of verifying interpolative results against (past) observations. The modification for backward T-FRI is proven to preserve many crucial properties that the original T-FRI possesses, e.g., the capability in handling multiple multi-antecedent rules and the maintenance of convexity and normality of interpolated results. Whilst backward T-FRI helps address the problem of missing antecedent attribute values, it does not totally remove this problem, especially when the scale of missing values becomes substantial.

- Dynamic T-FRI [Naik et al., 2017b]

A great majority of the existing transformation based FRI mechanisms work on a static sparse rule base. However, the use of a static rule base may affect the effectiveness of FRI due to the absence of the most concurrent (dynamic) rules as the requirements of fuzzy systems may change over time. Yet, a volume of intermediate fuzzy rules are typically generated from this type of FRI methods while executing rule interpolation. Collectively, they may gradually cover regions that were uncovered by the original sparse rule base, thereby offering possibly valuable information for updating the static sparse rule base. From this observation, the work of [Naik et al., 2017b] makes use of such intermediate rules which are otherwise discarded once the required outcomes have been obtained within the most of transformation based FRI methods, to develop a dynamic T-FRI mechanism. It enriches the rule base by refining and promoting the intermediate rules, gaining efficiency by allowing for more direct rule-firing while minimising future running of the interpolation procedure. It is implemented by first partitioning the interpolated rules into hypercubes, where the nonempty ones are fed as the input into a Genetic Algorithm-based clustering algorithm to find the "best" cluster arrangement. An iterative process is then run to select the densest clusters that have accumulated a sufficient number of candidate rules for achieving the rule aggregation and promotion. The practical significance of this approach is obvious. Further reinforcement is however, still possible, say, by employing more effective and efficient clustering and optimisation methods to replace the relevant components within the current implementation.

- Higher order T-FRI [Chen et al., 2016, Chen and Shen, 2017]

A common implementation shared by most of the established FRI methods is that fuzzy membership functions in the rules and observations are generally defined with conventional type-1 fuzzy sets, for the interpretation and treatment of uncertainty. Very little of the existing FRI work can conjunctively handle more than one form of uncertainty in the rules or observations, despite there may be cases in which more complicated fuzzy set representations become necessary [Fu and Shen, 2010]. In response to this observation, a higher order FRI has been developed, allowing for the representation and manipulation of different types of uncertainty in knowledge within the common T-FRI framework. It works by first encoding uncertain knowledge with higher order representation and then, by deriving the final conclusions through performing higher order interpolation over models of such representation. In particular, two common types of technique for uncertainty representation are exploited, resulting in: 1) a rough-fuzzy set based rule interpolation approach [Chen et al., 2016], which facilitates the representation of uncertain fuzzy membership functions with rough-fuzzy approximations; and 2) an interval type-2 fuzzy set-based approach [Chen and Shen, 2017], which works in the same manner as with the rough-fuzzy-based T-FRI. Within either method, the concept of representative value of a fuzzy set also plays an indispensable role as within the original type-1 T-FRI. Another type-2 fuzzy set-based FRI method can be found in [Chen and Lee, 2011]. Such methods require relative modifications corresponding to each particular uncertainty representation, which inevitably increases the computational complexity as the cost for exchange of a much more general T-FRI mechanism that will collapse back to the type-1 method if all uncertainty involved can be sufficiently captured by type-1 fuzzy sets.

- Other T-FRI-like approaches

Apart from the above-outlined modifications to T-FRI that are directly investigated and improved upon the original T-FRI method, there are other proposals for reinforcing fuzzy interpolative reasoning which are analogous to the basic ideas of T-FRI [Li et al., 2005]. For instance, in [Chen and Adam, 2018], ranking values of arbitrary polygonal fuzzy sets are used to express the characteristic points of the underlying membership functions, which are in turn used to play a similar role in the modified transformation-based FRI process as the *Rep* values do in T-FRI. In addition, the scale and move transformations involved in T-FRI are replaced with the distance ratio and move rate, respectively, to transform the constructed intermediate rule in an effort to obtain the final interpolated outcome.

Another variation of T-FRI is reported in [Chen and Ko, 2008], named CK FRI hereafter to acknowledge it inventors. The classical *Rep* values are substituted by characteristic values in this work, facilitating not only the simplified representation of a fuzzy set but also the definition of the distance between fuzzy terms. For situations where polygonal fuzzy sets are involved, the interpolated fuzzy set being sought is derived by calculating each of the characteristic points that are obtained from a series of α-cuts. Particularly, the normal points (of which the membership is 1) are first determined, aiding in any subsequent calculation of the remaining points. From this, two transformations, namely increment and ratio transformations, are executed to convert the average consequent into the final interpolated outcome with the similarity degree measured between these two analogous to that of the average antecedent and the observation. Improved on this work further, two enhanced transformations have been introduced [Chen et al., 2009] to support weighted approaches to FRI (that will be addressed separately later in this chapter).

2.2.3.4 Generalised Function-based FRI

Bearing significant similarity with the intermediate rule based FRI methods as outlined above is another approach, which is herein referred to as generalised function-based for convenience. Example methods falling within this family include those reported by [Baranyi et al., 1995, Baranyi et al., 1996a, Baranyi and Kóczy, 1996a, Baranyi et al., 2004, Baranyi et al., 1996b, Baranyi et al., 1998, Baranyi and Kóczy, 1996b]. Unlike the α-cut based interpolation algorithms, given an unmatched observation, this approach infers the conclusion based on the interpolation of fuzzy relations instead of using α-cut distances. It works through two major steps which are briefly outlined below for academic completeness. Further details can be referred to each relevant reference provided.

Given two fuzzy rules (say, r^1, r^2) and an observation (o^*) in the form of Eqn. (2.1), the core of a generalised function-based method can be described through the following two stages.

A. Generation of Interpolated Firing Rule

The aim of this first step is to create an intermediate rule r', in such a way that the antecedent of r' is as "close" to that of the observation (A^*) as possible. Note that the term "close" here stands for the case where at least partial overlapping is ensured between the observation and the intermediate rule. This implies the firing of the resulting intermediate rule can be subsequently conducted (see the next stage). Denote the procedure of this stage by

$$r' = f^{Interpolation}(r^1, r^2) \tag{2.37}$$

where $f^{Interpolation}$ represents a mapping from a pair of rules onto a set of all possible rules of the form as per Eqn. (2.1). There are two types of algorithm that may be utilised to implement this stage of the approach:

- Fuzzy relation interpolation, which includes any of the solid cutting methods [Baranyi et al., 1995, Baranyi et al., 1996a, Baranyi and Kóczy, 1996a, Baranyi et al., 1996b, Baranyi and Kóczy, 1996b], and those based on the fixed point law or the fixed value law [Ding et al., 1989, Ding et al., 1992, Mukaidono et al., 1990, Shen et al., 1993, Shen et al., 1988].
- Semantic relation interpolation, which includes any of the semantic revision methods [Ding et al., 1989, Ding et al., 1992, Mukaidono et al., 1990, Shen et al., 1993, Shen et al., 1988], using the semantic revision principle to describe the relation between the antecedents and consequent fuzzy sets within an interpolated intermediate rule.

B. Inference with Single Rule Firing

The second stage is to fire the intermediate rule returned by the first. This is enabled by temporarily regarding the newly generated intermediate rule as one of the existing rules within the rule base, and also by computing the overlapping between the observation and the antecedent of the intermediate rule. The procedure implementing this stage can be generally denoted by

$$B^* = f^{Inference}(r', A^*) \tag{2.38}$$

Exactly what mechanism to implement rule-firing may vary with respect to different FRI methods in this family. Any method reported in [Ding et al., 1989, Ding et al., 1992, Mukaidono et al., 1990, Shen et al., 1993, Shen et al., 1988] may be utilised to directly fire the transformed intermediate rule to compute the final consequent value required.

For simplicity, the above description has been focussed on interpolation for the cases where a single rule antecedent is considered. As with α-cut based and T-FRI approaches, the generalised function-based mechanism has also been extended to performing FRI with multiple rule antecedents and fuzzy extrapolation. More details can be found in [Baranyi et al., 2004] and other derivatives. Overall, the workflow of such a method can be illustrated in Fig. 2.5. Conceptually, this is of course very similar to the underlying approach taken by T-FRI.

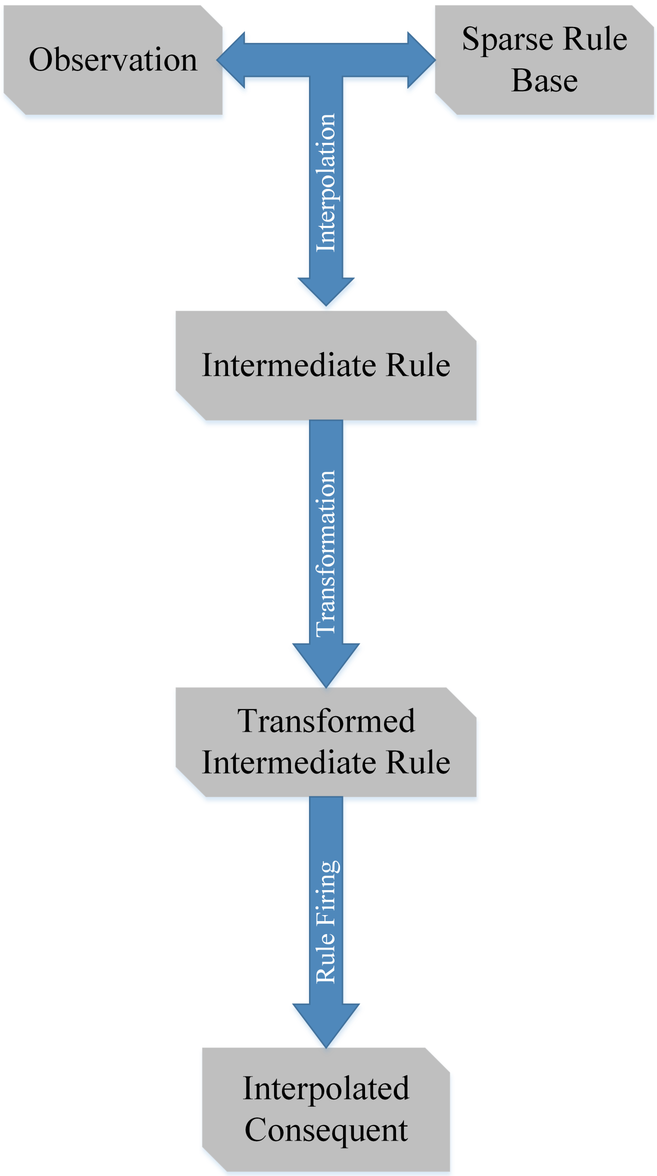


Figure 2.5: Workflow of generalised function-based fuzzy rule interpolation.

2.2.4 Comparison of Representative FRI Methods over Common Criteria

From the above, running an FRI algorithm, be it α-cut based or intermediate rule based, results in an inference consequent in response to an unmatched observation through interpolating the fuzzy rules in a given rule base, achieving the goal of interpolative reasoning. Theoretically, FRI is essentially a mapping [Tikk et al., 2011,Jenei, 2001] that relates the input space $\mathscr{A}$ and the output space $\mathscr{Z}$, where fuzzy subsets in the domain $\mathscr{A}$ and $\mathscr{Z}$ (denoted by $\mathscr{F}(\mathscr{A})$ and $\mathscr{F}(\mathscr{Z})$, respectively) indicate the values of rule antecedent attributes and the value of rule consequent (as defined in Eqn. (2.1)). That is, given a rule base R, $\forall r^i \in R$, the values $\{A_1^i, A_2^i, \ldots, A_m^i\} \in \mathscr{F}(\mathscr{A})$, of the m antecedent variables, and the rule consequent value $B^i \in \mathscr{F}(\mathscr{Z})$. FRI pursues to define the correlation $I : \mathscr{F}(\mathscr{A}) \rightarrow \mathscr{F}(\mathscr{Z})$, which associates to an observation $A^*(= \{A_1^*, A_2^*, \ldots, A_m^*\}) \in \mathscr{F}(\mathscr{A})$ an interpolated conclusion $I(A^*) = B^*$ where $B^* \in \mathscr{F}(\mathscr{Z})$. Thus, FRI methods are required to satisfy certain common properties as a mapping function, which also form the general criteria facilitating the comparison amongst them.

Table 2.6 summarises the most commonly used criteria for FRI evaluation in the literature. Any given FRI method is expected to meet or qualify at least a certain number of such properties to be effective in performing interpolative reasoning. Over the history of FRI development, a number of approaches that have been reported at the early stages have been analysed and compared against these criteria in the previous work of [Johanyák and Kovács, 2006, Tikk et al., 2011], especially for the α-cut based FRI methods including the seminal linear interpolation mechanism introduced in [Kóczy and Hirota, 1993a,Kóczy and Hirota, 1993b] and its derivations. Such comparative discussion is therefore, not comprehensively included in the present review to avoid redundancy. Instead, particular attention is drawn for more recently developed FRI approaches, including many popular transformation-based techniques. As a summary, Table 2.7 presents the results of evaluating the representatives of such FRI methods, over the common criteria.

In general, it is not necessary that all such criteria are fulfilled in developing an FRI method. However, it is expected that most of the property should be satisfied, with other problem specific parameters to fulfill given a certain application. This also points out the trend in the development of FRI techniques, that is to amend the

Table 2.6: Commonly Adopted Criteria for FRI Evaluation

No.	Criterion	Interpretation
C1	Validity of conclusion	Interpolated conclusion always leads to a valid fuzzy set.
C2	Preservation of convexity and normality	If an observation is normal and convex so should the interpolated conclusion also be.
C3	Compatibility with rule base	For all rules $r^i \in R$ and all $A^* \in \mathscr{F}(\mathscr{A})$: If $A^* = A^i$, then $\mathscr{I}(\mathscr{A}^*) = B^* = B^i$.
C4	Continuity condition	For $\epsilon > 0$ there exists $\delta > 0$ s.t. if $A, A^* \in \mathscr{F}(\mathscr{A})$, and $d(A, A^*) \leq \delta$ then $d(\mathscr{I}(A), \mathscr{I}(A^*)) \leq \epsilon$, where $d(\cdot,\cdot)$ denotes a certain distance metric.
C5	Preservation of piece-wise linearity	If fuzzy sets used for interpolation are piece-wise linear, so should interpolated conclusion be.
C6	Preservation of "in between"	If A^* is in between A^i and A^j, then interpolated conclusion $\mathscr{I}(A^*)$ should be in between $\mathscr{I}(A^i)$ and $\mathscr{I}(A^j)$.
C7	Use of arbitrary membership function	Mapping $\mathscr{I}$ is applicable to any convex and normal form of membership functions.
C8	Shape invariance	If all fuzzy sets in rule antecedents are of same type of membership function, so should interpolated $\mathscr{I}(A^*)$ be.
C9	Applicability for multiple rules	Interpolation mapping $\mathscr{I}$ can handle fuzzy interpolative reasoning with any number of rules.
C10	Applicability for multidimensional input	Interpolation mapping $\mathscr{I}$ is applicable to any finite number of input variables.
C11	Capability of extension to extrapolation	Interpolation mapping $\mathscr{I}$ is extrapolatively extensible.

drawbacks of the existing FRI methods and to satisfy more criteria. For example, the very first proposal for FRI, KH linear interpolation [Kóczy and Hirota, 1993a], is well-known that it cannot always guarantee the convexity of the derived fuzzy sets (i.e., C2 in Table 2.6) although they may be normal. This has led to much attention being paid to building FRI mechanisms that ensure not only normality but also convexity of inferred consequences. This in turn, has led many advanced variations of KH method. A number of recently developed FRI approaches are able to accomplish many key criteria successfully, including the listed C1-C7 as shown in Table 2.7. Also, criteria C9, C10 and C11 have increasingly become more demanded as more sophisticated fuzzy systems are constructed that enjoy more significant interpolative reasoning power.

2.3 Weighted Fuzzy Interpolative Reasoning

In the conventional fuzzy interpolative reasoning systems, multiple rules are generally involved with each concerning multiple rule antecedent attributes. However, these antecedent attributes are assumed to have equal significance when they are working together within the rule interpolation process. Recently, a number of methods have been proposed to weight the rule antecedents and to integrate the weights into the traditional algorithms where attributes are not weighted.

This section reviews the relevant weighted fuzzy interpolative reasoning systems in the literature. Table 2.8 lists the titles of the methods being reviewed, and an acronym is assigned to each to act as the short name after its inventors while reflecting the year of the relevant publication. The rest of this section is organised by first summarising four particular approaches, and then by presenting a brief comparison amongst them.

2.3.1 Typical Weighted FRI Approaches

This section reviews four representative fuzzy interpolative reasoning mechanisms which are achieved by weighted FRI. As indicated previously, two key issues, namely weight learning and integration of weights in FRI, are the main concerns in implementing any weighted FRI approach. The following subsections are therefore composed of three parts for each method, by first reporting the development regarding these two issues and then drawing summarising remarks. To facilitate better understanding, all weighted FRI methods are outlined by the use of unified pseudo code for the main procedural steps.

Table 2.7: Evaluation of Representative FRI Methods over Common Criteria

FRI Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
KH FRI	×	×	√	×	√	√	×	√(†)	√	√	√
							(† only if C1 is satisfied)				
CCL FRI	√	√	-	-	-	-	√	√	√	√	-
T-FRI	√	√	√	√	√	√	√	√	√	√	√
Adaptive T-FRI	√	√	√	√	√(†)	√	√	√	×	√	×
				(† first-order piecewise linear)							
Backward T-FRI	√	√	√	√	√	√	√	√	√	√	√
CK FRI	√	√	√	√	√	√	√	√	√	√	-
Dynamic T-FRI	√	√	√	√	√	√	√	√	√	√	√
Generalised function-based FRI	√	√	√	√	√(†)	√	√(⋆)	-	-	√	√
				(† when involving triangular fuzzy sets)							
					(⋆ for certain combinations of methods in this family)						

√ denotes satisfaction of certain criterion
× denotes dissatisfaction of certain criterion
- denotes information unavailable to decide on certain criterion

Table 2.8: Weighted Fuzzy Interpolative Reasoning Schemes with Short Names

Short Name	Title	Reference
LHTZ2005	Weighted fuzzy interpolative reasoning method	[Li et al., 2005]
CC2008	A new method for multiple fuzzy rules interpolation with weighted antecedent variables	[Chang and Chen, 2008]
CKCP2009	Weighted fuzzy interpolative reasoning based on weighted increment transformation and weighted ratio transformation techniques	[Chen et al., 2009]
CC2011a	Weighted fuzzy rule interpolation based on GA-based weight-learning techniques	[Chen and Chang, 2011b]
CC2011b	Weighted fuzzy interpolative reasoning for sparse fuzzy rule-based systems	[Chen and Chang, 2011a]
CLS2013	Weighted fuzzy interpolative reasoning systems based on interval type-2 fuzzy sets	[Chen et al., 2013b]
CH2014	Weighted fuzzy interpolative reasoning based on the slopes of fuzzy sets and particle swarm optimization techniques	[Chen and Hsin, 2014]
DJS2014	Antecedent selection in fuzzy rule interpolation using feature selection techniques	[Diao et al., 2014]
CC2016	Weighted fuzzy interpolative reasoning for sparse fuzzy rule-based systems based on piecewise fuzzy entropies of fuzzy sets	[Chen and Chen, 2016]
CA2018	Weighted fuzzy interpolated reasoning based on ranking values of polygonal fuzzy sets and new scale and move transformation techniques	[Chen and Adam, 2018]

2.3.1.1 Center of Gravity-based Weighted FRI (LHTZ2005)

A weighted FRI is presented in [Li et al., 2005] as the original approach for fuzzy interpolative reasoning supported with antecedent weights. This is referred to as the LHTZ2005 method in Table 2.8 and hereafter. All weights in this approach are implemented by the use of trapezoidal fuzzy sets.

A. Learning weights

There is little learning involved in LHTZ2005, but the weights of the individual rule antecedents are predefined by domain experts. Each antecedent attribute within different rules of the rule base is assigned a different weight. For instance, two weighted fuzzy rules used for weighted interpolation are therefore, represented such that [Li et al., 2005]:

$$\begin{aligned}
r^1 \; : \; if \;\; a_1 \; is \; A_1^1 \; (AW_1^1) \; and \; a_2 \; is \; A_2^1 \; (AW_2^1) \; and \cdots \; and \; a_m \; is \; A_m^1 \; (AW_m^1) \\
then \; z \; is \; B^1(C_1) \\
r^2 \; : \; if \;\; a_1 \; is \; A_1^2 \; (AW_1^2) \; and \; a_2 \; is \; A_2^2 \; (AW_2^2) \; and \cdots \; and \; a_m \; is \; A_m^2 \; (AW_m^2) \\
then \; z \; is \; B^2(C_2)
\end{aligned} \tag{2.39}$$

where AW_j^i stands for the weight of the jth antecedent variable $(j = 1, 2, \ldots, m)$ in the rule $r^i, i = 1, 2$, and $C_i, i = 1, 2$ is the certainty factor of r^i. Note that all of the AW_j^i and C_i are specified as trapezoidal fuzzy numbers. As such, the computational effort involved may generally increase significantly.

B. Weighting FRI

This weighted fuzzy interpolative reasoning process essentially extends the FRI mechanism of [Huang and Shen, 2003] that uses only triangular fuzzy sets. In this work, the center of gravity (COG) of a fuzzy set is used to represent the fuzzy set for simplicity. In particular, the COG of a trapezoidal fuzzy set $A = (a, b, c, d)$ is defined as a pair (h_L, h_R):

$$h_L = \frac{1}{3}(a + b + d) \qquad h_R = \frac{1}{3}(a + c + d) \tag{2.40}$$

where a, b, c, d denote the characteristic points of A with a and d having a membership of 0, and b and c being the odd normal points (i.e., the two extrema points of the nuclear of the trapezoidal with a membership of 1).

The distance between two trapezoidal fuzzy numbers A_1 and A_2 are defined using their COG pairs (namely, (h_{1L}, h_{1R}) and (h_{2L}, h_{2R})) as follows:

$$d(A_1, A_2) = \frac{1}{2}(h_{2R} + h_{2L} - h_{1R} - h_{1L}) \tag{2.41}$$

From this, LHTZ2005 can be summarised in Alg. 1, demonstrating the main execution steps. Note that the weights of rule antecedents are typefaced in bold wherever they are integrated within FRI in order to highlight the weighting mechanism.

Algorithm 1 Weighted FRI in LHTZ2005

Input:

- Rules r^1, r^2 in form of Eqn. (2.39)
- Observation $o^* = \{A_1^*, \ldots, A_m^*\}$, over m conditional attributes
- Individual weight, of rule antecedent variable $AW_j^i, i = 1, 2, j = 1, 2, \ldots, m$
- Certainty factors $C_i, i = 1, 2$ of rules $r^i, i = 1, 2$

Output:

- Interpolated consequent B^*

1: Construct a new inference rule by manipulating two given adjacent rules r^1, r^2, in form of

$$r' \ : \ if \ a_1 \ is \ A'_1 \ and \ a_2 \ is \ A'_2 \ and \cdots \ and \ a_m \ is \ A'_m, \ then \ z \ is \ B'$$

where $A'_j (j = 1, 2, \ldots, m)$ is obtained by

$$A'_j = (1 - \lambda_j) A_j^1 + \lambda_j A_j^2$$

with λ_j calculated using distance between fuzzy sets as per Eqn. (2.41):

$$\lambda_j = \frac{d(A_j^1 \otimes \mathbf{AW_j^1} \otimes C_1, A_j^*)}{d(A_j^1 \otimes \mathbf{AW_j^1} \otimes C_1, A_j^2 \otimes \mathbf{AW_j^2} \otimes C_2)}$$

where $\otimes$ denotes multiplication operator between trapezoidal fuzzy sets. B' is calculated similarly, such that

$$B' = (1 - \lambda_c) B^1 + \lambda_c B^2, \qquad \lambda_c = \frac{1}{m} \sum_{j=1}^{m} \lambda_j$$

2: Calculate scale rate s_j and move rate l_j $(j = 1, 2, \ldots, m)$ for each rule antecedent to assess difference between A'_j and A_j^*.

3: Aggregate scale and move rate for consequent variable by

$$s_c = \frac{1}{m} \sum_{j=1}^{m} s_j, \qquad l_c = \frac{1}{m} \sum_{j=1}^{m} l_j$$

4: Calculate consequent fuzzy set B^* by transforming B' using s_c, l_c. Note that antecedent weights are not involved here (more details can be referred to original work of [Li et al., 2005]).

5: **Return** Interpolated consequent B^*

C. Remarks

1. The weights of rule antecedents in this approach are assumed to be predefined, which requires human intervention and hence adversely reduces the flexibility of the resulting fuzzy interpolative reasoning system.

2. As reflected in Alg. 1, the individual weights for antecedent variables are only involved in the calculation of the aggregation factors $\lambda_j, j = 1, 2$ while constructing the new inference rule r' (as shown in Line 1). The aggregation over the scale and move rates to compute the consequent variable is simply implemented by an algebraic average of the corresponding antecedent items (Line 3), which are externally assigned, to signify their individual significance levels in influencing the consequent.

3. Computational complexity may be increased significantly due to the use of trapezoidal fuzzy sets to represent the weights, but the interpretability may be improved if these weights are associated with domain-specific linguistic terms (which is possible given they are defined by the domain experts).

2.3.1.2 GA-based Weighted FRI (CC2011a)

The method of genetic algorithm (GA)-based weighted fuzzy interpolative reasoning method integrates the weights of rule antecedents within the underlying FRI procedure it adopts. This work is referred to as CC2011a with details given in [Chen and Chang, 2011b]. In this method, the weights of the antecedent variables are automatically learned by the use of a GA-based weight-learning algorithm. The fuzzy sets are represented with polygonal or bell-shaped membership functions.

A. Learning weights

The learning method for generating the optimal weights of the rule antecedent variables used for this weighted FRI is based on the CHC algorithm [Eshelman, 1991], which is a specialisation of traditional GAs [Holland, 1975]. This GA-based learning mechanism encodes the weights of individual antecedents into a chromosome, on which each gene represents an individual attribute weight.

An initial population is randomly generated, forming the start point of the evolutionary weight learning process. For each chromosome in the current population, it decodes a certain weight value, which is to be employed in the proposed weighted FRI. The weighted interpolative scheme is then triggered for a set of training samples, with the interpolated outcomes recorded. The selection of "good" chromosomes depends on a predefined fitness function by comparing the values between the inferred outputs and the target outputs of the training samples. From this, a crossover operation is carried out among the selected chromosomes, forming the next generation. Once the number of evolutions reaches a predefined maximum number of evolutions, this iterative weight learning procedure terminates and the chromosome with the largest fitness value is deemed the optimal. The final learned weights for the rule antecedent variables are obtained by decoding the optimal chromosome. In so doing, this GA-based weight learning scheme follows a so-called "wrapper" approach, which mixes up weight learning and weighted FRI procedures. The weights obtained from the current generation are required to be integrated within FRI, to enable the evaluation of the fitness values.

B. Weighting FRI

A key concept employed for facilitating this weighted interpolative reasoning is the ratio of fuzziness (RF) [Chen and Chang, 2011b]. For polygonal fuzzy sets, the degree of fuzziness is computed in relation to the areas of the fuzzy sets. Let A and B be two polygonal fuzzy sets, the ratio of fuzziness $RF(A,B)$ of A to B is defined as follows:

$$RF(A,B) = \frac{S(A)}{S(B)} \tag{2.42}$$

where $S(A)$ and $S(B)$ denote the area of the membership function of A and that of B, respectively.

The central idea for the computation of the interpolated consequent in response to an (unmatched) observation is the following: The algorithm attempts to keep the RF of the fuzzy set of each attribute in the observation over that of the corresponding antecedent of a selected rule for interpolation the same as the RF of the fuzzy set of the required consequent (to be computed) over that of an artificially constructed intermediate rule consequent. The intermediate rule consequent is herein generated by aggregating the consequents of rules that are involved for interpolation. This idea in effect reflects fuzzy or generalised modus ponens, a keen to the approach taken by T-FRI.

The weights of the rule antecedent variables are integrated within FRI by following the routine which is summarised in Alg. 2. In order to emphasise the role that those antecedent weights play in the entire weighted FRI procedure, the individual weights are highlighted in bold in this algorithm description.

C. Remarks

1. The GA-based weight learning scheme requires many predefined parameters, such as fitness function and the maximum iteration number.
2. In the evolutionary learning process, the updating of weights requires the repeated runs of weighted FRI to compute the consequent using the current weights, in order to evaluate their fitness. This means the weight learning process is affected by the implementation of the underlying FRI process itself.
3. The individual weights of rule antecedents are only involved in the aggregation to obtain the rule weights, as shown in Line 3. They are not integrated with the underlying FRI.

2.3.1.3 Piecewise Fuzzy Entropies-based Weighted FRI (CC2016)

In [Chen and Chen, 2016], another method for weighted fuzzy interpolative reasoning is proposed through the exploitation of the concept of piecewise fuzzy entropies. This is referred to as CC2016 hereafter, which can handle fuzzy sets defined by polygonal and bell-shaped membership functions. In this method, weights are assigned differently to each antecedent variable when dealing with the same variable that is involved in different rules used for interpolation.

A. Learning weights

In CC2016 a fuzzy rule is generally represented in the following:

$$r^i\ :\ if\ a_1\ is\ A_1^i\ (AW_1^i)\ and\ a_2\ is\ A_2^i\ (AW_2^i)\ and \cdots\ and\ a_m\ is\ A_m^i\ (AW_m^i),\ then\ z\ is\ B^i$$

where AW_j^i stands for the weight for jth antecedent variable in the rule r^i. As indicated above, for a certain antecedent variable, its weight is allowed to be different

Algorithm 2 Weighted FRI in CC2011a

Input:

- Rule base $R = \{r^1, \ldots, r^N\}$ consisting of N rules in form of Eqn. (2.1)
- Observation $o^* = \{A^*_1, \ldots, A^*_m\}$, over m conditional attributes
- Number of closest rules n

Output:

- Interpolated consequent B^*

1: Initialise randomly individual weights of rule antecedent variables $AW_j, j = 1, 2, \ldots, m$;

2: Calculate representative values for antecedent fuzzy sets $A_j, j = 1, \ldots, m$ of rules $r^i, i = 1, \ldots, N$ and o^*, namely $Rep(A^i_j)$ and $Rep(A^*_j)$;

3: Select n nearest neighbouring fuzzy rules with respect to o^* from rule base;

4: Calculate weight for each of selected rules $W_i, i = 1, \ldots, n$, by aggregating **individual antecedent weights $\mathbf{AW_j}$**, $j = 1, 2, \ldots, m$ such that

$$W_i = \frac{Min_{1 \le j \le m} \mathbf{AW_j} \lambda_{ij}}{\sum_{q=1}^{n} Min_{1 \le j \le m} \mathbf{AW_j} \lambda_{qj}}$$

where λ_{ij} measures similarity between jth antecedent value of rule r^i and its corresponding value in o^*, using their representative values;

5: Construct (triangular) intermediate consequence fuzzy set $B' = (b'_1, b'_2, b'_3)$ (though any other polygonal MF may be used);

6: Divide membership function of each $A^i_j, i = 1, \ldots, n$ and $A^*_j, j = 1, \ldots, m$ into two (triangular) areas, namely $S_L(A)$ and $S_R(A)$, expressing such area on left hand side and that on right hand side of normal point, respectively;

7: Calculate RF of composite fuzziness of observed fuzzy sets $S(A^*_1, A^*_2, \ldots, A^*_m)$ over composite fuzziness of n weighted neighbouring antecedent fuzzy sets for each left and right area, such that

$$RF = \frac{S(A^*_1, A^*_2, \ldots, A^*_m)}{\sum_{i=1}^{n} W_i S(A^i_1, A^i_2, \ldots, A^i_m)}$$

$$S(A^*_1, A^*_2, \ldots, A^*_m) = \sum_{j=1}^{m} S(A^*_j) = \sum_{j=1}^{m} (S_L(A^*_j) + S_R(A^*_j))$$

$$S(A^i_1, A^i_2, \ldots, A^i_m) = \sum_{j=1}^{m} S(A^i_j) = \sum_{j=1}^{m} (S_L(A^i_j) + S_R(A^i_j))$$

8: Calculate characteristic points of interpolated consequent fuzzy set $B^* = (b^*_1, b^*_2, b^*_3)$, such that

$$b^*_2 = b'_2$$
$$b^*_1 = b^*_2 - RF \times \left| b'_2 - b'_1 \right|$$
$$b^*_3 = b^*_2 + RF \times \left| b'_3 - b'_2 \right|$$

9: **Return** Interpolated consequent $B^* = (b^*_1, b^*_2, b^*_3)$

when it is involved in different fuzzy rules. Such weights of rule antecedent attributes are generated during the weighted FRI process itself, which is explained next.

B. Weighting FRI

The fuzzy sets in this work are assumed to be polygonal, which are represented by their characteristic points (CPs), paired with the corresponding membership degrees [Chen and Chen, 2016]. Let A be a polygonal fuzzy set in the universe of discourse and the number of CPs for characterising A be n, then $A = (a_0, a_1, \ldots, a_l, a_c, a_r, \ldots, a_{n-1}; \mu_0, \mu_1, \ldots, \mu_{n-1})$, where a_l and a_r are called the "left normal point" and the "right normal point", and $a_c = \frac{a_l + a_r}{2}$ the "central point", with $\mu_0 = \mu_{n-1} = 0, \mu_l = \mu_r = 1$.

The basic idea is to construct an interpolated consequent fuzzy set B^* with regard to an input observation, by estimating its n CPs and the corresponding membership values such that $B^* = (b_0^*, b_1^*, \ldots, b_{n-1}^*; \mu_0^*, \mu_1^*, \ldots, \mu_{n-1}^*)$. The key step to perform the estimation of the membership degrees is carried out through computing the piecewise fuzzy entropies of the fuzzy sets involved. The concept of piecewise fuzzy entropy is defined via the notion of non-probability fuzzy entropy of a fuzzy set [Al-Sharhan et al., 2001, De Luca and Termini, 1972]. In particular, the piecewise fuzzy entropy $H_{t-1,t}(A)$ between the $(t-1)$th CP and the tth CP of a polygonal fuzzy set A is specified as below:

$$H_{t-1,t}(A) = -K \sum_{s=t-1}^{t} [\mu_s log_{10}(\mu_s) + (1-\mu_s) log_{10}(1-\mu_s)] \tag{2.43}$$

where $K = 1/n$ and $1 \leq t \leq n-1$, and μ_s denotes the degree of membership of the CP a_s.

The weighted interpolative approach is summarised in Alg. 3, with the individual weights highlighted in bold where they are learned and used.

C. Remarks

1. As indicated above, in this method, the weights for individual rule antecedent variables are assigned differently when different rules involving them are taken into consideration.

Algorithm 3 Weighted FRI in CC2016

Input:

- Rule base $R = \{r^1, \ldots, r^N\}$ consisting of N rules in form of Eqn. (2.1)
- Observation $o^* = \{A_1^*, \ldots, A_m^*\}$, over m conditional attributes

Output:

- Interpolated consequent B^*

—Estimation of CPs of B^*

1: Calculate representative values for antecedent fuzzy sets $A_j, j = 1, \ldots, m$ of rules $r^i, i = 1, \ldots, N$ and observation o^*, namely $Rep(A_j^i)$ and $Rep(A_j^*)$;

2: Calculate relative placement factor γ_j for A_j^* with respect to antecedent fuzzy sets of its neighbouring rules using their corresponding *Rep* values, and obtain mean value of these:

$$\gamma = \frac{\sum_{j=1}^{m} \gamma_j}{m}, \; j = 1, 2, \ldots, m$$

and

$$\gamma_j = \frac{Rep(A_j^*) - Rep(A_{lj})}{Rep(A_{rj}) - Rep(A_{lj})}$$

where A_{lj} and A_{rj} are left and right closest fuzzy set of A_j^*, respectively, in the ascending order of $A_j^i, i = 1, \ldots, N, *$.

3: Calculate CPs of B^* such that

$$b_t^* = (1 - \gamma) \times b_t^l + \gamma \times b_t^r$$

where b_t^*, b_t^l, b_t^r are tth CP of B^*, B^l, B^r, respectively, and $1 \le t \le n-1$, n denotes number of CPs used for characterising a polygonal fuzzy set, $l = \lfloor \frac{(n-1)}{2} \rfloor$ and $r = \lceil \frac{(n-1)}{2} \rceil$.

—Estimation of membership values with respect to CPs of B^*

4: Calculate piecewise fuzzy entropy between $(t-1)$th CP and tth CP for each A_j^i, A_j^* and B^i following Eqn. (2.43), namely $H_{t-1,t}(A_j^i)$, $H_{t-1,t}(A_j^*)$ and $H_{t-1,t}(B^i)$ where $i = 1, 2, \ldots, N, j = 1, 2, \ldots, m$ and $1 \le t \le n-1$.

5: Calculate central point $a_{j,c}^i, a_{j,c}^*, b_{j,c}$ of each A_j^i, A_j^* and B^i, respectively where $i = 1, 2, \ldots, N, j = 1, 2, \ldots, m$.

6: Calculate **weight** $\mathbf{AW_j^i}$ for jth antecedent variable in rule r^i as follows:

$$\mathbf{AW_j^i} = 1 - \frac{d(A_j^*, A_j^i)}{max_{1 \le h,k \le N} d(A_j^h, A_j^k)}$$

(see next page for continuation)

(continuation of Weighted FRI in CC2016)

where $d(A_j^*, A_j^i)$ and $d(A_j^h, A_j^k)$ are defined by

$$d(A_j^*, A_j^i) = \left| a_{j,c}^* - a_{j,c}^i \right|, \quad d(A_j^h, A_j^k) = \left| a_{j,c}^h - a_{j,c}^k \right|$$

with $1 \leq j \leq m, i, h, k \in 1, 2, \ldots, N$.

7: Calculate weight for each rule $W_i, i = 1, \ldots, N$, by aggregating **antecedent weights** $\mathbf{AW_j^i}$, $j = 1, 2, \ldots, m$ such that

$$W_i = \frac{Min_{j=1,\ldots,m}\mathbf{AW_j^i}}{\sum_{i=1}^{N} Min_{j=1,\ldots,m}\mathbf{AW_j^i}}$$

8: Calculate piecewise fuzzy entropy $H_{t-1,t}(B^*)$ between $(t-1)$th CP and tth CP of B^*:

$$H_{t-1,t}(B^*) = \begin{cases} \sum_{j=1}^{m} \left\{ H_{t-1,t}(A_j^*) \times \left[\sum_{i=1}^{N} W_i \times \frac{H_{t-1,t}(B^i)}{\sum_{j=1}^{m} H_{t-1,t}(A_j^i)} \right] \right\} if \ \exists i, j H_{t-1,t}(A_j^i) > 0 \\ \sum_{j=1}^{m} H_{t-1,t}(A_j^*) \quad if \ \forall i, j \ H_{t-1,t}(A_j^i) = 0 \end{cases}$$

where $1 \leq t \leq n-1$.

9: Calculate membership degrees $\mu_0^*, \mu_1^*, \ldots, \mu_{n-1}^*$ with regards CPs of interpolated result B^*, using piecewise fuzzy entropy $H_{t-1,t}(B^*), 1 \leq t \leq n-1$ as per method of [Chapra and Canale, 1998] (more details of which can be found in [Chen and Chen, 2016]).

10: Obtain interpolated fuzzy set B^* in terms of CPs (from Line 3) and corresponding membership values (from Line 9), resulting in

$$B^* = (b_0^*, b_1^*, \ldots, b_{n-1}^*; \mu_0^*, \mu_1^*, \ldots, \mu_{n-1}^*)$$

11: **Return** Interpolated consequent B^*

2. The generation of the antecedent weights is achieved during the weighted FRI process, as shown in Line 5 of the algorithm.

3. The individual weights of rule antecedents are only involved in the aggregation stage to obtain the overall rule weights, as shown in Line 6. Unfortunately, such useful information is not integrated within the rest of the fuzzy interpolative reasoning process.

2.3.1.4 Weighted Increment and Ratio Transformation-based Weighted FRI (CKCP2009)

Another weighted FRI method is presented in [Chen et al., 2009], referred to as CKCP2009 hereafter. It uses weighted increment transformation and weighted ratio transformation to enable weighted fuzzy interpolative reasoning. A "wrapper" algorithm is implemented for automatically tuning the optimal weights of the antecedent variables appearing in a fuzzy rule, capable of dealing with polygonal, Gaussian and bell-shaped membership functions.

A. Learning weights

The weights on individual rule antecedent variables are automatically learned within a "wrapper" mechanism. The weighted interpolation process is required to be iteratively triggered in order to update the current weights. Particularly, the weight learning procedure within the proposed weighted FRI is tailored for a certain system control problem, where one input may lead to several states indicating the current values of the observation. The weight learning process is summarised below.

Individual weights are initialised with the same value to start the first iteration. A set of training samples as rule antecedent attribute values are then employed as the input to the FRI system, together with the current weights, resulting in the next states of these variables. To adjust the weighting value of each rule antecedent attribute, the gradient-descent training method is utilised, where a predefined fitness function over the rule antecedent variable is evaluated using the value recorded in its final state. The fitness function generates the prediction error for each antecedent variable in the current iteration and the weights are then modified with the aim to minimise the error, which are subsequently employed to run the next iteration. The entire iterative weight updating process is terminated when a preset maximum number of iterations is reached.

B. Weighting FRI

As with many FRI methods reviewed previously using *Rep* values, a unique real value is also defined herein and associated with a certain fuzzy set for reflecting the key information on the overall location in its domain. In CKCP2009, for a polygonal fuzzy set $A = (a_1, a_2, \ldots, a_{n-1})$, the characteristic value $CV(A)$ (or *Rep* as termed elsewhere) is defined as follows:

$$CV(A) = \frac{(a_0 + a_1 + \cdots + a_{n-1})}{n} \tag{2.44}$$

The distance between two polygonal fuzzy sets P and Q is then specified by the use of their CV values such that

$$d(P, Q) = \left| CV(P) - CV(Q) \right| \tag{2.45}$$

Given these notions, this weighted FRI method can be summarised as shown in Alg. 4 [Chen et al., 2009], where the multiplication operation between a polygonal fuzzy set A and a real value w ($w \in [0, 1]$) is defined by

$$A \otimes w = (a_1, a_2, \ldots, a_{n-1}) \otimes w = (wa_1, wa_2, \ldots, wa_{n-1}) \tag{2.46}$$

As with the other weighted FRI approaches, the individual weights are highlighted in bold within the algorithm description.

C. Remarks

1. This weight learning scheme is an iterative process. It is integrated within the weighted interpolation procedure, of which the outcome is required to be collected to evaluate the fitness function to update the current weights.

2. Although the weights are designed to be automatically tuned for optimisation, the approach is tailored to a specific problem, where the fitness functions for each rule antecedent are predefined. This limits the generality of the underlying techniques.

3. This algorithm reflects the intuition in approximate reasoning in that "how an observation is transformed from an intermediate antecedent fuzzy set should be reflected in how the interpolated outcome is transformed from the intermediate consequent". This is basically the same as the idea adopted by the conventional T-FRI.

Algorithm 4 Weighted FRI in CKCP2009

Input:

- Rules $R = \{r^1, r^2\}$ in form of Eqn. (2.39)
- Observation $o^* = \{A_1^*, \ldots, A_m^*\}$, over m conditional attributes

Output:

- Interpolated consequent B^*

1: Initialise individual weights of rule antecedent variable $AW_j^i, i = 1,2, j = 1,2,\ldots,m$;

2: Construct fuzzy set $A'_j = (a'_{j1}, a'_{j2}, a'_{j3})$ from fuzzy sets A_j^1 and A_j^2 such that Let

$$a'_{j1} = a^*_{j2} - la'_{j,1-2}$$
$$a'_{j2} = a^*_{j2}$$
$$a'_{j3} = a^*_{j2} + la'_{j,2-3}$$

where

$$la'_{j,1-2} = (1-\lambda^j_{rep})la^1_{j,1-2} + \lambda^j_{rep} la^2_{j,1-2}$$
$$la'_{j,2-3} = (1-\lambda^j_{rep})la^1_{j,2-3} + \lambda^j_{rep} la^2_{j,2-3}$$

$$la^i_{j,1-2} = d(a^i_{j1}, a^i_{j2}) \qquad la^i_{j,2-3} = d(a^i_{j2}, a^i_{j3}) \quad i = 1,2. \;\; j = 1,2,\ldots,m$$

and

$$\lambda^j_{rep} = \frac{d(A^1_j \otimes \mathbf{AW^1_j}, A^*_j \otimes (\mathbf{AW^1_j} + \mathbf{AW^2_j})/2)}{d(A^1_j \otimes \mathbf{AW^1_j}, A^2_j \otimes \mathbf{AW^2_j})}$$

3: Calculate normal point (i.e, b^*_2) of interpolated fuzzy set B^* such that

$$b^*_2 = (1-\lambda)CV(B^1) + \lambda CV(B^2)$$

where

$$\lambda = \sum_{j=1}^{m} \frac{\mathbf{AW^1_j} + \mathbf{AW^2_j}}{W} \lambda_j \qquad W = \sum_{j=1}^{m} (\mathbf{AW^1_j} + \mathbf{AW^2_j})$$

$$\lambda_j = \frac{d(A^1_j, a^*_j)}{d(A^1_j, A^2_j)} = \frac{\left|CV(A^1_j) - a^*_j\right|}{\left|CV(A^1_j) - CV(A^2_j)\right|}$$

(see next page for continuation)

(continuation of Weighted FRI in CKCP2009)

4: Construct fuzzy set $B' = (b'_1, b'_2, b'_3)$ based on fuzzy sets B^1 and B^2 as with that shown in Line 2, such that

$$\begin{aligned} b'_1 &= b^*_2 - lb'_{1-2} \\ b'_2 &= b^*_2 \\ b'_3 &= b^*_2 + lb'_{2-3} \end{aligned}$$

where

$$\begin{aligned} lb'_{1-2} &= (1-\lambda_{rep})lb^1_{1-2} + \lambda_{rep} lb^2_{1-2} \\ lb'_{2-3} &= (1-\lambda_{rep})lb^1_{2-3} + \lambda_{rep} lb^2_{2-3} \end{aligned}$$

with

$$lb^i_{1-2} = d(b^i_1, b^i_2) \qquad lb^i_{2-3} = d(b^i_2, b^i_3) \quad i = 1, 2$$

$$\lambda_{rep} = \sum_{j=1}^{m} \left(\frac{\mathbf{AW^1_j} + \mathbf{AW^2_j}}{\sum_{j=1}^{m} (\mathbf{AW^1_j} + \mathbf{AW^2_j})} \right) \lambda^j_{rep}$$

5: Calculate extreme points (namely b^*_1 and b^*_3) of interpolated fuzzy set B^* as follows: Given artificially constructed antecedent and consequent fuzzy sets $A'_j, j = 1, 2, \ldots, m$ and B', by aggregating corresponding components within two given rules, analogical reasoning is performed to derive remaining characteristic points of interpolated result B^*, i.e., b^*_1 and b^*_3. This is achieved through two subroutines, namely increment and ratio transformations, which are implemented by assessing similarity between each of A'_j and A^*_j first and then, applying similarity measure to B' in order to obtain B^*. **Antecedent weights AW** are computed as shown in Line 3 (more computational details are omitted but can be found in [Chen et al., 2009]);

6: **Return** Interpolated consequent $B^* = (b^*_1, b^*_2, b^*_3)$

2.3.2 Comparison of Existing Weighted FRI Methods

The above four weighted fuzzy interpolative reasoning mechanisms are typical approaches from the viewpoint of weight learning and FRI weighting. This subsection contrasts these approaches and categorises other weighted FRI methods listed in Table 2.8 in relation to the distinct features of these four approaches.

2.3.2.1 Weight Learning Mechanisms

As reflected by the preceding subsections, typical approaches to weighted fuzzy interpolative reasoning contain basic properties along with which other weighted FRI methods can be grouped and compared.

A. Predefined vs. Automatically learned

The initial idea for obtaining weights on rule antecedent attributes is simply to predefine them with domain expertise directly acquired from the experts. This approach includes the early work as reported in LHTZ2005 and CC2008. It requires human intervention and hence, adversely reduces the flexibility of the resulting fuzzy systems. Automatic weight learning schemes are obviously preferred. Indeed, all of the remaining methods in Table 2.8 pursue alternative ways to learn weights automatically.

B. Unique weight vs. Multiple weights for an antecedent attribute

In general, weighted FRI works with fuzzy rule bases that involve multiple rule antecedent variables. Different significance levels are associated with different variables to indicate their different contributions towards the conclusion. In the literature, for a given rule antecedent attribute, certain methods learn a unique weight for each variable independent of what rules that variable appears in, whilst others assign different weights to one common attribute in different rules. The former includes work in CC2011a, CC2008, CH2014, DJS2014 and CC2011b, and the latter includes LHTZ2005, CC2016, CKCP2009 and CA2018. When a rule antecedent attribute may be assigned with multiple weights, depending on which rules it may appear in, the overall rule model becomes more complicated and harder to interpret. Moreover, more specific information regarding the antecedent variables of observations may become necessary, in order to compute the characteristic points of the corresponding fuzzy sets. This may include for example, information on central point [Chen and Chen, 2016] or that on ranking value [Cheng et al., 2015] of the fuzzy set, thereby at the expense of involving more computation to produce the weights than otherwise. Besides, in so doing, the weights are only measurable during the running of the weighted FRI system when an observation is provided.

C. Filter schemes vs. Wrapper schemes

The terms *filter* and *wrapper* are used to group the weight learning schemes, based on their dependence upon whether a weighted FRI method will be recursively called on during the process of weighted generation. That is, these weight learning methods following the filter scheme are independent of the weighted FRI process, whereas the wrapper methods need to exploit the outcome of the weighted FRI in order to evaluate the "goodness" or quality of the current weights. The filter approach is taken by CC2016, DJS2014 and CA2018, and the wrapper approach by CC2011a, CKCP2009, CH2014 and CC2011b. Since methods belonging to the wrapper group employ interpolated results for constructing fitness functions (in an effort to update the required weights in the current iteration), their performance in terms of accuracy may be very high, but the computational overheads is relatively costly at the same time.

2.3.2.2 Weighting FRI Procedures

This issue is concerned with how the generated weights of rule antecedent attributes are integrated within the underlying FRI, for revealing the relative significance level of each individual attribute in contributing to the derivation of the interpolated results. As can be seen from the typical weighted FRI mechanisms reviewed previously, the following observations can be drawn:

1. Existing techniques generally work by artificially creating an overall weight to each of the rules before running the weighted rules in FRI. Such weights are normally computed through aggregating the weights calculated for individual rule antecedent variables, thereby involving additional weight aggregation procedures.

2. Learned weights are seldom systematically integrated within all major components of the weighted FRI algorithm, but just involved in certain computational subroutines. As such, information regarding domain attribute significance is not exploited to its full potential. Also, different underlying FRI mechanisms employ the weights in different manners, limiting the generality of these approaches and their transplantability to suit other FRI methods.

There are also several weighted fuzzy interpolative reasoning schemes that reflect other perspectives. For example, a particular work in [Chen et al., 2013b] constructs a weighted FRI method based on interval type-2 fuzzy sets. However, these viewpoints are beyond the scope of this thesis and hence, their details are omitted.

In summary, the above review of weighted fuzzy interpolative reasoning confirms the importance of developing techniques that allow for differentiating the relative significance levels of individual domain attributes. In particular, weights should be learned automatically and efficiently, ideally without requiring additional observations or triggering the entire FRI system. Also, it is desirable to create a general weighting scheme that enables different non-weighted FRI methods to be supported with antecedent weights in a common manner. In so doing, it helps facilitate transplanting a developed weighting scheme from one FRI mechanism to another once the weights of rule antecedent attributes are available.

2.4 Attribute Evaluation within Feature Selection

Feature selection (FS) aims to choose a minimal subset of domain features (interchangeably termed attributes or variables hereafter) that are the most relevant to the target concept or decision. As the resultant features are directly selected from an original feature set, the FS techniques preserve the original meaning of the selected attributes while reducing the dimensionality of the original set.

As indicated previously (in Chapter 1), FS is achieved within a four-step procedure. Amongst them, an evaluation function is used to measure how good an attribute is, or a subset of attributes are, regarding the potential solution to the problem at hand. This offers a natural way to evaluate the relative significance of an attribute. If systematically carried out across all domain attributes, the use of such a function will enable the ranking of the attributes with regard to the underlying quality criteria.

Existing evaluation functions in the literature can be generally grouped into categories that reflect the different criteria adopted to judge attribute quality, including those based on measures over distance, information, dependence, consistency, etc [Dash and Liu, 1997]. From the perspective of how these attributes are evaluated and selected, two major classes can be found for attribute evaluation within FS. One is to assess each feature individually, resulting in the most informative features that are ranked on the top being selected; the other is based on feature subsets, where a subset of features are measured jointly. In the following subsections, a brief introduction is outlined to these two groups of methods, covering those that are popularly used and readily available.

2.4.1 Individual Feature Ranking-based Methods

This subsection presents four attribute evaluation schemes which are based on assessing and ranking each feature individually.

2.4.1.1 Information Gain (IG)

Information gain (IG) has been widely adopted in the development of learning classifier algorithms, to measure how well a given attribute may separate the training examples according to the underlying classes [Mitchell, 1997]. It is defined via the entropy metric in information theory [Shannon, 2001], which is commonly used to characterise the disorder or uncertainty of a system.

Formally, let $\mathbf{O} = (O, p)$ be a discrete probability space, where $O = \{o_1, o_2, \ldots, o_n\}$ is a finite set of domain objects, with each having the probability $p_i, i = 1, \ldots, n$. Then, the Shannon entropy of O is defined by

$$Entropy(O) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.47}$$

Regarding the task of classification, $o_i, i \in \{1, \ldots, n\}$ represents a certain object, and p_i is the proportion of O which is labelled as the class $j, j = 1, \ldots, m, m \leq n$. Note that the entropy is at its minimum (i.e., $Entropy(O) = 0$) if all elements of O belong to the same class (with $0 \log_2 0 = 0$ defined); and the entropy reaches its peak point (i.e., $Entropy(O) = \log_2 n$) if the probability of o_i belonging to each category is equal; otherwise the entropy is between 0 and $\log_2 n$.

Intuitively, the less the entropy value, the easier the classification problem. It is based on this observation that information gain has been introduced to measure the expected reduction in entropy caused by partitioning the values of an attribute. This leads to the popular decision tree learning methods [Quinlan, 1986]. Given a collection of examples $U = \{O,A\}$, $o_i \in O$ $(i = 1,\dots,n)$ is an object which is represented with a group of attribute $A = \{a_1,\dots,a_l\}$ and a class label m. Information gain upon a particular attribute $a_k, k \in \{1,\dots,l\}$, is defined as

$$IG(O,a_k) = Entropy(O) - \sum_{v \in Value(a_k)} \frac{|O_v|}{|O|} Entropy(O_v) \tag{2.48}$$

where $Value(a_k)$ is the set of all possible values for the attribute a_k, O_v is the subset of O where the value of the attribute a_k is equal to v (i.e., $O_v = \{o \in O | a_k(o) = v\}$), and $|\cdot|$ denotes the cardinality of a set.

From the perspective of entropy evaluation over U, the second part of Eqn. (2.48) shows that the entropy is measured via weighted entropies that are calculated over the partition of O using the attribute a_k. The bigger the value of information gain $IG(O,a_k)$, the better the partitioning of the given examples with regards a_k. Obtaining a high information gain therefore, implies achieving a significant reduction of entropy or uncertainty caused by considering the influence of that attribute.

2.4.1.2 Relief-F

Relief-F [Kononenko, 1994] is an extension of the original Relief method [Kira and Rendell, 1992b, Kira and Rendell, 1992a] developed for estimating attribute significance efficiently. It can be used to deal with noisy, incomplete and multi-class data sets, working by exploiting distance measures. Each individual attribute is assigned a cumulative weight computed over a predefined number of sample data selected from a given training data set. Attributes with a weight above a certain threshold become selected elements of the attribute subset sought.

A weight is assigned on the basis of the following intuition: Instances that belong to a similar class should be closer together than those in a different class. Suppose that $near\_hit$ represents an instance that is closest to a certain training instance x under consideration, with both belonging to the same class, and that $near\_miss$ represents an instance that is closest to x but in a different class. The cumulative weight associated with a given attribute is then computed by

$$w_i = w_{i-1} - d(x, near\_hit)^2 + d(x, near\_miss)^2, i = 1, \ldots, I \tag{2.49}$$

where $w_0 = 0$, I stands for the number of training iterations, and $d(.,.)$ is typically implemented with Euclidean metric.

2.4.1.3 Laplacian Score (LS)

Laplacian score (LS) [He et al., 2006] is another distance measure-based attribute evaluation function. It is calculated for each individual attribute to reflect its capability of locality preserving. The definition of LS is inspired by an observation that the data points being related to the same topic should be close to each other.

Let LS_k denote the LS measure of a certain attribute a_k. It is computed by

$$LS_k = \frac{\sum_{ij}(f_{ki} - f_{kj})^2 S_{ij}}{Var(f_k)} \tag{2.50}$$

where f_{ki} and f_{kj} denote the value of a_k within the instance x_i and that within x_j, respectively, $Var(f_k)$ is the estimated variance of a_k, and S_{ij} represents the neighbourhood relationship between the instances x_i and x_j, such that

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}}, \text{ if } x_i \text{ and } x_j \text{ are nearest neighbours} \\ 0, \text{ otherwise} \end{cases} \tag{2.51}$$

A quality attribute should be of a small Laplacian score.

2.4.1.4 Local Learning-based Clustering for FS (LLCFS)

LLCFS [Zeng and Cheung, 2010] performs attribute selection within the framework of Local Learning-based Clustering (LLC) [Wu and Schölkopf, 2007]. It computes a weight and assigns it to each attribute while performing a clustering task. Typically, the weights are thinly distributed if the dataset contains much redundancy, with a weight of zero indicating that the corresponding attribute is dispensable; only those attributes associated with a weight of a significant magnitude are selected.

LLCFS works by iteratively executing the following two steps until convergence: (i) estimating the weights for the attributes using intermediate clustering results, and (ii) updating the clustering with weighted attributes. As such, the weights are estimated iteratively during the clustering process.

Incidentally, such an FS approach is termed wrapper-based in the literature, as opposed to the other techniques outlined herein which follow the so-called filter-based approach to FS [Liu and Motoda, 2012]. The filter-based and wrapper-based schemes are specified in terms of their dependence on the inductive algorithm that will finally use the selected subset. Filter methods are independent of the inductive algorithm, whereas wrapper approaches employ the inductive algorithm as the evaluation function. This is similar in concept to the description of "wrapper" and "filter" approaches to learning attribute weights as presented in Section 2.3.

2.4.2 Feature Subset Evaluation-based Methods

Another four attribute evaluation schemes that follow the idea of assessing feature subsets (instead of individual attributes) are reviewed below.

2.4.2.1 Rough Set-based Feature Selection (RSFS)

Rough set-based FS (RSFS) [Chouchoulas and Shen, 2001, Jensen and Shen, 2004, Shen and Chouchoulas, 2002] employs the concepts of rough set theory [Pawlak, 1982, Pawlak, 1996, Pawlak, 2012] to distinguish the significance of attributes. Let $I = (U, A)$ be an information system, where U is a non-empty set of finite objects (the universe of discourse) and A is a non-empty finite set of attributes. An equivalence relation associated with a particular subset $P \in A$ is $IND(P)$:

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\}$$

This relation generates the partition of U, which is denoted as $U/IND(P)$:

$$U/IND(P) = \otimes\{a \in P : U/IND(\{a\})\}$$

where

$$M \otimes N = \{X \cap Y : \forall X \in M, \forall Y \in N, X \cap Y \neq \oslash\}$$

Note that objects in the partition of U generated by the relation $IND(P)$ have the same values over all attributes in subset P. In other words, such objects are *indiscernible* by attributes from P.

Let the equivalence classes of the P-indiscernibility relation be denoted by $[x]_P$. Rough set approximates a subset of universe of discourse $X \subseteq U$ by utilising a pair of sets in approximation space, namely *P-lower approximation* $P_*(X)$ and *P-upper approximation* $P^*(X)$, respectively:

$$P_*(X) = \bigcup_{x \in U} \{[x]_P : [x]_P \subseteq X\}$$
$$P^*(X) = \bigcup_{x \in U} \{[x]_P : [x]_P \cap X \neq \oslash\}$$

From an intuitive prospect, the *P-lower approximation* of X is the set of all objects, which can be for *certain* classified as X with respect to P, and the *P-upper approximation* of X is the set of all objects, which can be *possibly* classified in X in view of P.

Let P and Q be two subsets of attributes, and their associated equivalence relations over U be $IND(P)$ and $IND(Q)$ respectively. Then, the notion positive region is defined as:

$$POS_P(Q) = \bigcup_{X \in U/IND(Q)} P_*(X)$$

It contains all objects of U that can be classified as the classes of $U/IND(Q)$ using the information about the attributes P. Hence, the dependency of Q on P can be defined by

$$\gamma_P(Q) = \frac{\left|POS_P(Q)\right|}{|U|} \tag{2.52}$$

where $|\cdot|$ denotes the cardinality of set. Obviously, $\gamma_P(Q) \in [0, 1]$ is held. The closer it is to 1, the more Q depends on P.

Inspired by observation on the dependency factor, a measurement of significance of attribute subset can be defined by the reduction of dependency of Q on a set of attributes P while removing a particular one of them. That is, given P, Q and an attribute $a \in P$,

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q) \tag{2.53}$$

The larger the reduction of dependency, the more significant the attribute a is. This gives the manner which is to be used as a heuristic for FS.

RSFS works via generating possible subsets of attributes and selecting the one which leads to the maximum rough set dependency degree. The generation procedure can be achieved by pruning techniques to reduce the time complexity. The maximum rough set dependency should be equal to the dependency on all conditional attributes (if the dataset given is consistent). For FS, the cardinality of the selected feature subset should of course, be as small as possible. More formally, suppose that D denotes a decision attribute and C denotes all conditional attributes. Then, the searched reduct, which is denoted by R_{min}, can be computed such that

$$\begin{aligned} R_{min} &= \{X : X \in R, \forall Y \in R, |X| \leq |Y|\} \\ R &= \{X : X \subseteq C, \gamma_X(D) = \gamma_C(D)\} \end{aligned} \tag{2.54}$$

where RSFS is a purely data-driven approach selecting attributes from the data available. However, it works for situations involving discrete values only. To handle real valued features, fuzzy-rough feature selection is developed as outlined below.

2.4.2.2 Fuzzy-Rough Feature Selection (FRFS)

Fuzzy-rough set based feature selection (FRFS) approach [Jensen and Shen, 2004, Jensen and Shen, 2007, Jensen and Shen, 2008, Jensen and Shen, 2009] is used to discover data dependencies and to reduce the number of attributes contained within a data set using the data alone without acquisition of additional information. This method employs fuzzy positive region based on the concepts of fuzzy-rough sets to define a feature subset dependency function that assesses the relative significance of a certain subset of features. As with the dependency function in RSFS, this fuzzy-rough dependency function may also be used to evaluate the degree of importance of each individual feature if a subset contains just one element.

Fuzzy rough sets employ similar notions to those used in rough sets. By introducing fuzzy concepts into rough sets as an extension of rough set-based approach, the *fuzzy lower* and *fuzzy upper* approximation are defined as follows:

$$\mu_{P_*(X)} = \sup_{F \in U/IND(P)} min(\mu_F(x), \inf_{y \in U} max\{1 - \mu_F(y), \mu_X(y)\})$$
$$\mu_{P^*(X)} = \sup_{F \in U/IND(P)} min(\mu_F(x), \sup_{y \in U} min\{\mu_F(y), \mu_X(y)\})$$

Compared with RSFS, FRFS defines the dependency measurement by employing the fuzzy lower approximation. In particular, the membership function of the fuzzy positive region is defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X \in U/IND(Q)} \mu_{P_*(X)}(x) \tag{2.55}$$

Thus, the new dependency function now is:

$$\gamma_P(Q) = \frac{\left|\mu_{POS_P(Q)}(x)\right|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_P(Q)}(x)}{|U|} \tag{2.56}$$

where $\sum_{x \in U} \mu_{POS_P(Q)}(x)$ denotes the fuzzy cardinality of fuzzy membership function $\mu_{POS_P(Q)}(x)$.

Based on the fuzzy-rough set-based dependency function, the significance of a certain attribute can be measured by the reduction of this dependency degree with respect to the attribute. Indeed, the resulting FS procedure performs in the same manner as RSFS. There is a slight difference between FRFS and RSFS in that (for a consistent dataset) RSFS searches for a smallest subset of attributes whose dependency degree is the same as the whole conditional attributes, whereas the searching procedure in FRFS terminates when the dependency will not increase by adding any one of remaining attributes.

Note that the basic idea of FRFS has been extended to producing a number of more advanced techniques [Jensen and Shen, 2009], including that employing fuzzy similarity relations. These help to reduce the computational complexity caused by calculating the Cartesian product of fuzzy equivalence classes. These are not purely data-driven techniques however, the fuzzy similarity relations need to be predefined prior to their applications.

2.4.2.3 Correlation-based Feature Selection (CFS)

As a filter type of FS algorithm, CFS [Cui et al., 2010,Hall, 2000] exploits a correlation based heuristic to evaluate the worth of features. In particular, it employs the correlation coefficients amongst features to construct an evaluation function. The core principle is formulated as quantitatively assessing the quality of a given feature subset S:

$$Q(S) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2.57}$$

where $\overline{r_{cf}}$ and $\overline{r_{ff}}$ denote the average, feature-class and feature-feature correlations, respectively, and k is the number of features contained within S. In fact, Eqn. (2.57) is Pearson's correlation where all attributes have been standardised [Ghiselli, 1964].

This reflects the core heuristic for CFS to evaluate the merit or significance of a subset of features: "*Good feature subsets contain features highly correlated with class, yet uncorrelated with each other.*" Following this heuristic, features which are highly correlated to the decision attribute but independent to each other are deemed to be more significant and are therefore, more likely to be chosen in the process of CFS.

2.4.2.4 Consistency-based Feature Selection (IRFS)

The consistency measure presented in [Dash and Liu, 2003] for FS is defined through the introduction of the concept of *inconsistency rate* (IR) over a certain training dataset for a given feature set. IR is calculated via a three-stage process:

i) Determining all patterns that are inconsistent, where a pattern (or a part of an instance without class label) is considered inconsistent if there exists at least two instances such that they match all but their class labels.

ii) Calculating the inconsistency count for each pattern of the feature subset, which is the number of times this pattern appears in the data minus the largest of such a number among different class labels.

iii) Computing the inconsistency rate of the feature subset that is defined by the sum of all the inconsistency counts over all patterns of that subset appearing in the data, divided by the total number of the training instances.

The selected feature subset is expected to gain the smallest summation of inconsistency rates. From the above listed IRFS procedures, the consistency measure is able to work when data has discrete valued features. Any continuous feature is supposed to be discretised firstly by the use of any discretisation method before it is dealt with IRFS. Incidentally, this methodology for handling continuous data may be equally applied to RSFS.

2.4.3 Discussion

As can be seen from the above, FS approaches within both categories, namely feature ranking-based and feature subset evaluation-based, facilitate a variety of ways to evaluate features in an individual or group manner. More specifically, FS methods that follow the feature ranking-based scheme directly weigh and hence, rank individual features, which may follow a filter or wrapper based approach. The feature subset evaluation-based methods evaluate the quality of an attribute subset, instead of that of an individual attribute. However, if the subset is set to contain just one feature at a time, then the group-based approach becomes directly applicable to individual feature ranking also. These different styles of FS mechanism are all considered here in order to demonstrate the generality of ranking and weighting conditional attributes in fuzzy rules for guiding non-weighted fuzzy interpolative reasoning with the weights of rule antecedents, which will be illustrated in detail in the following chapters.

Finally, note that the evaluation functions embedded within FS techniques can also be categorised into two different types in terms of whether or not the decision attribute is involved when evaluating the conditional features. This leads to the supervised attribute evaluation schemes that include methods such as IRFS [Dash and Liu, 2003] and Relief-F [Kononenko, 1994], and the unsupervised schemes that include those like LS [He et al., 2006] and LLCFS [Zeng and Cheung, 2010]. Unsupervised approach offers more flexibility since the consequent attribute is not required during the attribute evaluation process.

2.5 Summary

This chapter has systematically reviewed the relevant background knowledge that forms key foundations for the investigations carried out in this project, as to be reported in the subsequent chapters. In particular, the fundamental fuzzy rule interpolation (FRI) techniques have been comprehensively analysed, resulting in the division of two major groups of methodologies, namely, the α-cut based interpolations and the intermediate rule based interpolations. These two types of approach are distinguished mainly by whether a so-called intermediate rule is to be used to facilitate the inference. Representative FRI approaches have been introduced for each category. Specific attention has been paid to one of the most popular intermediate rule based interpolation methods, the scale and move transformation based FRI (T-FRI), which has the closest relation to this project.

In the recent literature for FRI, there has been a significant volume of effort to enhance the potential of conventional FRI mechanisms by considering the rule antecedent weights within the original unweighted FRI procedure. This has been inspired by the observation that a common assumption was shared for traditional FRI methods in that the rule antecedent attributes are of equal importance. This is obviously an impractical assumption for many real-world problems. Two crucial issues are identified for developing the weighted FRI, namely weight generation and weight integration. Whilst relevant literature has shown limited progress, much work needs to be done in order to establish a weighted fuzzy interpolative reasoning system that would work effectively and efficiently.

Also covered in the review are the mechanisms for attribute evaluation. Particularly addressed are those that are embedded in the feature selection techniques. This is because they are mature techniques and readily available, which can be easily modified to assess the significance of individual rule antecedent attributes, thereby facilitating the learning of the corresponding weights, as to be shown next.

Chapter 3

Weight Learning from Rule Bases

As indicated previously, the first key issue for achieving a weighted fuzzy interpolative reasoning mechanism is to evaluate the relative significance levels of rule antecedent attributes appearing in the given rule base. Several alternatives for generating weights of the rule antecedents have been reported in the literature (as reviewed in Chapter 2). From a computational viewpoint, to automate the weight generation process, it has an intuitive appeal to take a two-step approach: 1) collecting data for the evaluation of the rule antecedent attributes, and 2) looking for an appropriate and applicable technique to evaluate the attributes using the available data. Without requiring any further information other than the given sparse rule base for the problem at hand, the question is whether such a two-step procedure can still be implemented. That is, can weights of rule conditional attributes be generated by the use of the sparse rule bases only? This chapter addresses this important question.

The rest of this chapter is structured as follows. Section 3.1 introduces how the data useful for attribute evaluation can be (re-)produced from a given sparse rule base. Section 3.2 illustrates how the rule antecedent weights can be generated from the data available. To further illustrate how the basic theoretical mechanisms perform their work, a practical case study is provided, which is integrated within the above two routines. Finally, Section 3.3 summarises the chapter.

3.1 Data Generation: Reverse Engineering for Rule Base Sparseness Reduction

In conventional FRI algorithms, the first key stage is the selection of n closest fuzzy rules to an observation when the observation is presented with no matching rules available in the rule base. In such work, all rule antecedents are assumed to be of equal significance while searching for the subset of closest rules; there is no assessment regarding the relative importance or ranking of these antecedents. This may reflect a seemingly important observation where typically, the fuzzy rules that are provided by domain experts or learned from historical data (which constitute the rule base) are of the form as shown in Eqn. (2.1). That is, there is no information available on the relative significance of individual antecedent attributes. This is a premier reason that the most existing approaches to FRI commonly assume the use of this format of knowledge representation. In real-world problems however, it is often the case where different domain attributes are of different significance.

Fortunately, the evaluation functions embedded in the feature selection (FS) techniques offer an effective ranking mechanism to address this issue. Nevertheless, while utilising the evaluation function of a certain FS method to differentiate the significance levels of the domain attributes, data is required to act as the training instances for computing their relative ranking scores. Therefore, the biggest problem of learning weights in an effort to distinguish the relative significance degrees associated with the conditionals is where the data comes from.

In general, FRI works with a sparse rule base. This implies that in the first place, it may be difficult to acquire sufficient example data for use in support of computing the required weights. Had there been sufficient training data in the problem domain, the situation of having a sparse rule base might not have existed, as such data could have been utilised to generate a dense rule base. Thus, only the originally provided sparse rule base is used as the information source for assessing attribute weights. This requires the introduction of a method to preprocess the sparse rule base for the generation of a valid set of training instances.

An innovative ***reverse engineering*** procedure is proposed here to address this issue. This is doable because every FRI system has a sparse rule base consisting of rules as represented in the form of Eqn. (2.1). This set of rules can be translated into a man-made decision table, forming a set of artificially created training samples, where each row represents either a rule in the given rule base or an artificial rule generated from a given rule. Note that in data-driven learning, rules are learned from data samples. The work here is done through a reverse engineering process of data-driven learning, translating rules back to data.

3.1.1 Reverse Engineering Procedure

The basic idea of the reverse engineering procedure is to reformulate automatically the rules in the given sparse rule base into data representations of a common structure. This is necessary because for a sparse rule-based system, different conditional attributes may appear in different rules and different rules may have different numbers of conditionals. Reflecting this view, the training instances are artificially generated through the following three steps:

i) Identification of all conditional attributes appearing in all the rules and all (finite fuzzy) values used to define these attributes;

ii) Expansion of each rule in the sparse rule base into one that involves all conditional attributes such that if a certain conditional is not originally involved in the rule, then it is inserted into the rule with its value being set to a qualitative term, "don't care"; and

iii) Replacement of each "don't care" with every possible fuzzy value for the corresponding attribute in each rule that contains this qualitative value, such that one rule involving L attributes of the "don't care" value ($L \geq 1$) is replaced by $\prod_{i=1}^{L} c_i$ rules, with c_i being the cardinality of the value domain of a certain conditional that does not appear in the original rule.

In so doing, within each of the expanded rules a conditional attribute that does not appear in a given rule now takes one and only one possible fuzzy value from its underlying domain. For example, if a given original rule contains just one "missing" conditional attribute, then this rule is expanded to k rules where k is the number of the fuzzy sets that this attribute may take as its value.

This reverse engineering procedure can be logically justified: For a given rule in the sparse rule base, if an attribute is missing from the rule antecedent, then the rule will have the same consequent value independent of what fuzzy value that attribute may take, provided that all other attributes appearing in the rule are satisfied regarding their respectively specified value. The presumption of the value domains being finite and discrete is also justifiable given that only fuzzy rules are considered here, where each attribute takes values from a (normally small) collection of fuzzy sets. In particular, the proposed reverse engineering procedure works with a sparse rule base, which typically involves a much smaller number of rules than the usual fuzzy rule-based systems. Besides, only those missing antecedents are to be filled with the possible fuzzy sets taken from their value domains. These factors jointly help restrain the adverse impact of the curse of dimensionality possibly caused by converting individual rules in the sparse rule base into artificial training samples.

3.1.2 Illustration of Reverse Engineering

A simple artificial example may help illustrate the idea of this procedure. Suppose that the sparse rule base consists of the following two rules only, each involving one different antecedent attribute, x or y, and the common consequent attribute z:

r_1: if x is A_1, then z is C_1

r_2: if y is B_2, then z is C_2

where x takes values from the domain $\{A_1, A_2\}$ and y from $\{B_1, B_2, B_3\}$.

Following the three-step reverse engineering procedure, first, all possible antecedent attributes involved in the problem are identified, these are x and y, together with their value domains as indicated above. Then, a temporary artificial decision table is generated by inserting "don't care" as the values taken by those antecedent attributes which are missing in the original rules, as shown in Table 3.1. The final artificial decision table as of Table 3.2 can then be constructed.

This creation of the artificial table is implemented as such, because there are two antecedent attributes in question, of which x has two possible values (A_1 and A_2) and y has three alternatives (B_1, B_2, B_3). Without losing generality, suppose that the first given rule is used to construct part of the emerging artificial decision table

Table 3.1: Temporary Artificial Decision Table

Artificial Rules \ Variables	x	y	z
r_1	A_1	"don't care"	C_1
r_2	"don't care"	B_2	C_2

first. As y is missing in r_1, which means if x is satisfied (with the value A_1), this rule is satisfied and hence, the consequent attribute z will have the value C_1 no matter which value y takes. That is, r_1 can be expanded into three artificial rules, resulting in r^1, r^2 and r^3 in Table 3.2, for each of which y takes one of its three possible values. Similarly, r^4 and r^5 can be constructed to expand the original rule r_2.

Table 3.2: Example for Reverse Engineered Decision Table

Artificial Rules \ Variables	x	y	z
r^1	A_1	B_1	C_1
r^2	A_1	B_2	C_1
r^3	A_1	B_3	C_1
r^4	A_1	B_2	C_2
r^5	A_2	B_2	C_2

3.1.3 Inconsistency in Artificial Decision Table

When the reverse engineering procedure is applied to a given (sparse) rule base, the resultant, artificially constructed decision table may include logically inconsistent rules where certain rules may have the same antecedent but different consequents. For instance, in the above illustrative example, r^2 and r^4 in Table 3.2 may appear to be inconsistent. This does not matter as the eventual rule-based inference, including rule interpolation does not use these artificially generated rules, but the original sparse rule base. They are created purely to help assess the relevant significance of individual variables through the estimation of their respective ranking scores. It is because there are attributes which may lead to potentially inconsistent implications in a given problem that it is possible to distinguish their relative importance to the problem (or their potential power in influencing the derivation of the consequent).

3.1.4 Practical Illustrative Case Study — Case Description

To help illustrate the proposal in this thesis, for attribute weighted fuzzy interpolative reasoning, a case study is utilised while being integrated within individual steps of the entire inference procedure. This illustrative case includes a commonly used fuzzy classification problem [Yuan and Shaw, 1995], involving a small set of training data of 16 instances. The system is set to make a decision on what sports activity to be undertaken (namely, volleyball, swimming and weight lifting) given the status of four conditional attributes regarding the weather, in terms of temperature (hot, mild and cool), outlook (sunny, cloudy and rain), humidity (humid and normal) and wind (windy and not windy).

Six fuzzy rules have been generated as given below using a standard rule induction algorithm [Yuan and Shaw, 1995]. These six rules form a dense rule base where the domains of the antecedent variables are completely covered by the rules. To facilitate the illustration (of rule interpolation), *Rule 6* is purposefully removed to have a sparse rule base.

1. If *Temperature* is *Hot* and *Outlook* is *Sunny*, then *Swimming*.
2. If *Temperature* is *Hot* and *Outlook* is *Cloudy*, then *Swimming*.
3. If *Outlook* is *Rain*, then *Weight lifting*.
4. If *Temperature* is *Mild* and *Wind* is *Windy*, then *Weight lifting*.
5. If *Temperature* is *Mild* and *Wind* is *Not Windy*, then *Volleyball*.
6. (If *Temperature* is *Cool*, then *Weight lifting*.)

3.1.5 Illustrative Case Study — Stage 1: Reverse Engineering

From the earlier theoretical demonstration, the *reverse engineering* procedure, as a crucial tool, converts the rules into a set of artificial training samples given a rule base, forming a decision table for the subsequent calculation of required attribute weights. In this illustrative case, the rule base presented in Section 3.1.4 (bar *Rule*

Table 3.3: Rule Base in Illustrative Case

Rules \ Variables	Temperature	Outlook	Humidity	Wind	Decision
r^1	*Hot*	*Sunny*	-	-	*Swimming*
r^2	*Hot*	*Cloudy*	-	-	*Swimming*
r^3	-	*Rain*	-	-	*Weight lifting*
r^4	*Mild*	-	-	*Windy*	*Weight lifting*
r^5	*Mild*	-	-	*Not Windy*	*Volleyball*

6) is reformulated as given in Table 3.3. As can be seen, a few values of certain rule antecedents are missing.

Recall the three-step procedure presented in Section 3.1.1, 32 training data are generated as listed in Table 3.4. The reverse engineering process is explained using this illustrative case. Without losing generality, assume that the first given rule is used to create the artificial data first. Then, part of the emerging artificial decision table is first constructed from this rule. Note that *Humidity* and *Wind* are missing in *Rule 1*, which means if *Temperature* is satisfied with the value *Hot* and *Outlook* with *Sunny*, the rule is satisfied and thus, the consequent variable *Decision* will have the value of *Swimming* no matter which values *Humidity* and *Wind* may take. That is, *Rule 1* can be expanded by the first four data in Table 3.4, each having the variables *Humidity* and *Wind* taking one of their respective two possible values. Similarly, more artificial data can be created by translating and expanding the remaining original rules.

Examining both the antecedent and the consequent values in Table 3.4, it can be seen that there are several identical samples which are generated from different original rules. Retaining one and only one of the same samples results in a total of 30 training data. Note that in such an artificially constructed decision table, it may again appear to include inconsistent data since they may have the same values for the respective antecedent attributes but different consequents (e.g., two inconsistent pairs are highlighted in Table 3.4). This makes no difference to the eventual rule-based inference, including rule interpolation, as previously discussed in Section 3.1.3. However, this enables the measuring of the attribute weights of individual antecedent variables as to be described next.

Table 3.4: 32 Training Data after Reverse Engineering Given Five Rules

Temperature	Outlook	Humidity	Wind	Decision
Hot	Sunny	Humid	Windy	Swimming
Hot	Sunny	Humid	Not Windy	Swimming
Hot	Sunny	Normal	Windy	Swimming
Hot	Sunny	Normal	Not Windy	Swimming
Hot	Cloudy	Humid	Windy	Swimming
Hot	Cloudy	Humid	Not Windy	Swimming
Hot	Cloudy	Normal	Windy	Swimming
Hot	Cloudy	Normal	Not Windy	Swimming
Hot	Rain	Humid	Windy	Weight lifting
Hot	Rain	Humid	Not Windy	Weight lifting
Hot	Rain	Normal	Windy	Weight lifting
Hot	Rain	Normal	Not Windy	Weight lifting
Mild	Rain	Humid	Windy	Weight lifting
Mild	Rain	Humid	Not Windy	Weight lifting
Mild	Rain	Normal	Windy	Weight lifting
Mild	Rain	Normal	Not Windy	Weight lifting
Cool	Rain	Humid	Windy	Weight lifting
Cool	Rain	Humid	Not Windy	Weight lifting
Cool	Rain	Normal	Windy	Weight lifting
Cool	Rain	Normal	Not Windy	Weight lifting
Mild	Sunny	Humid	Windy	Weight lifting
Mild	Sunny	Normal	Windy	Weight lifting
Mild	Cloudy	Humid	Windy	Weight lifting
Mild	Cloudy	Normal	Windy	Weight lifting
Mild	Rain	Humid	Windy	Weight lifting
Mild	Rain	Normal	Windy	Weight lifting
Mild	Sunny	Humid	Not Windy	Volleyball
Mild	Sunny	Normal	Not Windy	Volleyball
Mild	Cloudy	Humid	Not Windy	Volleyball
Mild	Cloudy	Normal	Not Windy	Volleyball
Mild	Rain	Humid	Not Windy	Volleyball
Mild	Rain	Normal	Not Windy	Volleyball

3.2 Weights Evaluation: Feature Ranking for Weighting Rule Antecedent

Having gone through the reverse engineering procedure for a given sparse rule base, an artificial decision table is derived that can be employed as a set of training instances, from which the weights of the rule antecedent attributes are learned. Such a decision table is also termed as a training instance pool that is to be used interchangeably hereafter. The significance levels, or weights, of the rule antecedent attributes are then assessed using an appropriate attribute evaluation method. This is because weighting the rule antecedent attributes essentially is to evaluate the relative significance of each of them, which may therefore be implemented through ranking amongst the attributes.

As indicated previously, the evaluation functions embedded in FS techniques offer an effective ranking mechanism to accomplish this task. In general, two types of the evaluation scheme exist in the literature, where one is to assess each attribute individually and respectively, while the other is to perform feature subset-based evaluation, where a group of features are assessed jointly. As reviewed in Chapter 2, any of the different types of feature ranking method may be applied to evaluate the relative significance of individual antecedent attributes, which is demonstrated in the following.

3.2.1 Scoring Individual Attributes

The category of attribute evaluation functions reviewed in Section 2.4.1 can be directly applied to assess individual attributes, including Information Gain (IG), Relief-F, Laplacian Score (LS) and Local Learning-based Clustering for FS (LLCFS). Application of each results in a vector of weighting scores associated with those attributes assessed. For easy referencing, these score vectors are denoted as $Score_{IG}$, $Score_{Relief-F}$, $Score_{LS}$ and $Score_{LLCFS}$, respectively.

Note that the LS-based FS method seeks those attributes of the smallest Laplacian score(s) for selection. Thus, the ranking score of LS for a rule antecedent attribute $a_i, i = 1, \ldots, m$, can be defined by

$$Score_{LS}(a_i) = \frac{1}{1 + LS_i} \tag{3.1}$$

This is just one of the possible alternatives, other transformation for example, e^{-LS_i}, may be employed to define the LS-based scores.

The other category of FS methods summarised in Section 2.4.2 conducts the selection process based on evaluating feature subsets, instead of individual attributes however. To obtain individual feature scores using any of these techniques the evaluation procedure needs to be modified. The following presents the modified version for use in this work. Note that as for the present problem, all entries in the training instance pool are discrete values by nature, Fuzzy-Rough FS (FRFS) is not applied here (since it deals with real-valued problems).

3.2.1.1 Rough Set-based FS (RSFS) for Scoring

It is known that the dependency degree $\gamma_P(Q)$ captures the dependence of an attribute subset Q on another subset P. Suppose that the subset Q contains the single consequent attribute z and the subset P contains just one certain antecedent attribute $a_i, i = 1, \ldots, m$ of a rule in the sparse rule base. Then, the general form of the dependency degree $\gamma_{\{a_i\}}(\{z\})$ between two subsets of attributes as per Eqn. (2.52) degenerates to one that assesses the importance degree of each individual antecedent attribute upon which the consequent depends:

$$Score_{RSFS}(a_i) = \gamma_{\{a_i\}}(\{z\}) = \frac{\left|POS_{\{a_i\}}(\{z\})\right|}{|U|} \tag{3.2}$$

This is of course, what RSFS exactly does in the first round during its iterative process of adding attributes to the emerging selected feature subset (starting from an empty set), determining which attribute is individually speaking, the best to be selected. It means that to obtain attribute scoring vector using the evaluation function of RSFS, only one iteration of the FS algorithm is needed to be run.

3.2.1.2 Correlation-based FS (CFS) for Scoring

Let the number of features be equal to 1, i.e., $k = 1$ in Eqn. (2.57), the evaluation criterion used in CFS is then modified to

$$Score_{CFS}(a_i) = Q(\{a_i\}) = r_{\{z\}\{a_i\}} \tag{3.3}$$

The feature subset S is replaced by a single antecedent variable of a given rule, degenerating to such a simple formula where only the feature-class correlation remains. This simplification has an intuitive appeal since any degree of correlation between an individual antecedent variable and a possible consequent reflects the fact that there is a certain contribution made by the given antecedent towards that consequent.

3.2.1.3 Consistency-based FS (IRFS) for Scoring

The modification of the consistency measure for scoring individual antecedent variables lies on the modification of the inconsistency rate (IR). Similar to the last two methods, the feature subset consisting of a single antecedent is assessed, resulting in the summation of the IR of all patterns of this antecedent. Note that the more important an antecedent variable, the smaller the IR. Hence, the score of each antecedent $a_i, i = 1, \ldots, m$, can be calculated by

$$Score_{IRFS}(a_i) = \frac{1}{1 + \sum_{allpatterns} IR} \tag{3.4}$$

where the highest score indicates the most significant antecedent variable amongst all.

3.2.2 Attribute Weighting

Having computed the scores of individual attributes, using either of the aforementioned scoring methods, a normalised relative weighting scheme can be readily introduced. Thus, all antecedent attributes employed in the rules of a given sparse rule base can be ranked, each (say, the attribute a_i) being associated with a weight:

$$W_i = \frac{Score_*(a_i)}{\sum_{t=1,\ldots,m} Score_*(a_t)} \tag{3.5}$$

where $Score_*$ denotes any of the seven types of score (namely, one of the following: $Score_{IG}$, $Score_{Relief-F}$, $Score_{LS}$, $Score_{LLCFS}$, $Score_{RSFS}$, $Score_{CFS}$ and $Score_{IRFS}$).

Given their underlying definition, the resulting normalised values have a natural appeal to be interpreted as the relative significance degrees of the individual rule antecedent attributes, in the determination of the corresponding rule consequent. Therefore, they can be used to act as the weights associated with each individual antecedent attribute in the original sparse rule base. Of course, for any implementation in modifying conventional non-weighted FRI, one and just one of the seven types of the weight is required. From this viewpoint, this work presents a range of choices regarding the weighting methods that may be utilised to support and refine fuzzy interpolative reasoning, as to be described in next chapter.

3.2.3 Illustrative Case Study — Stage 2: Weights Generation

The proposed weights generation mechanism is herein illustrated by continuing the example of Section 3.1.4 and following the illustration in Section 3.1.5. As indicated previously, it is only one of the modified feature ranking methods that is required at a time for one implementation, although any of those methods available may be taken to assess the relative significance of individual antecedent attributes. Information gain (IG) is utilised here due to its popularity and the simplicity of demonstration. Others may be utilised in a similar manner.

In the following, the computational process regarding the weights generation by the use of IG is explicitly shown, for tailoring it to the illustrative case. Given an artificial decision table that is derived from a sparse rule base via reverse engineering, the information gain IG_i of a certain antecedent variable $a_i, i = 1, \ldots, m$, regarding the consequent variable z can be calculated, as per its definition in Eqn. (2.48) as follows:

$$IG_i = Entropy(\{z\}) - \sum_{v \in Value(a_i)} \frac{|\{z\}_v|}{|\{z\}|} Entropy(\{z\}_v) \tag{3.6}$$

where $\{z\}_v$ denotes the subset of rules in the artificial decision table in which the antecedent variable a_i has the value v. Repeating the above, the information gains for all antecedent variables $IG_i, i = 1, \ldots, m$ can be computed. These values are then normalised into the attribute weights $W_i, i = 1, \ldots, m$ according to the weights generation scheme described in Section 3.2.2, such that

$$W_i = \frac{IG_i}{\sum_{t=1,\ldots,m} IG_t} \tag{3.7}$$

For the example case, the normalised IGs calculated for each antecedent variable using those 30 training samples (reduced from Table 3.4 by eliminating identical data) are shown in Table 3.5. The weight of the antecedent attribute *Temperature* is relatively higher than those for the other three, which indicates *Temperature* plays a more important role in the decision on the sports activity. This can be verified from examining the five fuzzy rules, where the antecedent variable *Temperature* appears in four rules. On the other hand, *Humidity* and *Wind* are assigned a very small or zero-valued weight. In particular, the normalised weight of *Humidity* is 0, signifying its irrelevance on the decision in this rule base.

Table 3.5: Normalised Weights Calculated by Information Gains

Antecedent	*Temperature*	*Outlook*	*Humidity*	*Wind*
Normalised Weights	0.5000	0.4515	0.0000	0.0485

3.3 Summary

In this chapter, the first key issue for constructing a weighted FRI algorithm has been addressed, regarding the automatic generation of the weights of rule antecedent attributes. In particular, the data used for the weight learning is derived from the given sparse rule base only, through creating a training instance pool via a reverse engineering procedure. The individual weights are then evaluated by the use of the attribute ranking mechanism extracted and modified from a given FS technique. The resultant learned weights are uniquely associated with each of rule antecedent variables, no matter which rule it is involved in. Importantly, the entire weight learning scheme exploits just the knowledge already available, i.e., the sparse fuzzy rules in the given rule base, without acquisition of any observations nor that of any other information.

Chapter 4

Weighted Transformation-based Fuzzy Rule Interpolation

As illustrated in Chapter 3, the weights of individual rule antecedent attributes appearing in all rules of a given sparse fuzzy rule base can be computed through a reverse engineering procedure. From this, conventional fuzzy interpolative reasoning mechanism where all rule antecedents are unrealistically treated as of equal significance can be improved. For easy cross-referencing, the original FRI algorithms with no weighting imposed upon antecedent variables are interchangeably termed as non-weighted or unweighted methods hereafter. Correspondingly, those modified ones with weights are called weighted FRI methods.

This chapter presents an investigation into the issue of *how weights of rule antecedents are integrated within non-weighted FRI*. In general, FRI achieves interpolative inference given an observation that matches no rules in the given rule base, by the use of a small number of fuzzy rules which are closer to the observation (in other words, which have more similarity to the observation than the others). As such, the first step of FRI is normally the selection of what rules to be used to perform interpolation. These selected rules are interchangeably termed as the closest or the nearest neighbouring rules to the observation. The FRI is then executed with the selected closest rules regardless of the rest in the rule base. As an initial study of weighting conventional non-weighted FRI, the work is, herein, implemented by adapting the popular and award-winning scale and move transformation-based FRI (T-FRI) [Huang and Shen, 2006, Huang and Shen, 2008]. In particular, the weights are used to modify all components of the T-FRI computation process systematically, including the closest rule selection and rule interpolation.

This chapter is structured as follows. Sections 4.1 and 4.2 present the integration of antecedent weights within the closest rule selection and rule interpolation in T-FRI, respectively. In order to further demonstrate the theoretical work of the proposed weighted T-FRI, Section 4.3 illustrates the working procedure of the methods described in the first two sections, by continuing the illustrative example of Chapter 3. Section 4.4 discusses the key differences between the original non-weighted T-FRI method and the proposed attribute weighted modification. Section 4.5 describes the workflow of weighted T-FRI in implementation and a general framework for fuzzy rule-based inference that is supported by weighted FRI. Finally, Section 4.6 draws a summary of the work presented in the chapter.

4.1 Weighted Selection of Fuzzy Rules for Interpolation

Any FRI process starts as an observation o^* being newly presented to the fuzzy system does not activate any rule in the sparse rule base, due to no matching (or in certain FRI-based systems, due to too low level a partial matching). Thus, a neighbourhood of n ($n \geq 2$) closest rules of the observation is required to be chosen in order to perform rule interpolation. The conventional T-FRI approach to making this choice is by exploiting the Euclidean distance measured through aggregating the distances between individual antecedent attributes of a given rule and the corresponding attribute values in the observation, as per Eqn. (2.20). Now that the weights of individual attributes have been obtained with a scoring mechanism (derived from the use of the evaluation function in a feature selection method), the distance between a given rule r^p and the observation o^* needs to be updated accordingly.

Let $W = \{W_j | j \in 1, 2, \ldots, m\}$ be the collection of the weights on the antecedent attributes in the problem domain, with m denoting the number of all antecedent attributes concerned. Then, the traditional distance measure is modified such that

$$\begin{aligned} \tilde{d}(r^p, o^*, W) &= \frac{1}{\sqrt{\sum_{t=1}^{m}\left(\frac{1-W_t}{m-1}\right)^2}} \sqrt{\sum_{j=1}^{m}\left(\left(\frac{1-W_j}{m-1}\right) d(A_j^p, A_j^*)\right)^2} \\ &= \frac{1}{\sqrt{\sum_{t=1}^{m}(1-W_t)^2}} \sqrt{\sum_{j=1}^{m}\left((1-W_j) d(A_j^p, A_j^*)\right)^2} \end{aligned} \quad (4.1)$$

where $d(A_j^p, A_j^*)$ is calculated according to Eqn. (2.21) which is herein recited below for completion:

$$d(A_j^p, A_j^*) = \frac{\left| Rep(A_j^p) - Rep(A_j^*) \right|}{max_{A_j} - min_{A_j}} \tag{4.2}$$

Note that $d(A_j^p, A_j^*)$ represents the normalised result of the otherwise absolute distance; max_{A_j} and min_{A_j} denote the maximal and minimal value of the attribute a_j, respectively. As stated previously, triangular membership functions are used throughout the reasoning process, the representative value of a fuzzy set can be simply calculated by averaging the vertices of the triangular membership function, such that

$$Rep(A) = \frac{v_1 + v_2 + v_3}{3} \tag{4.3}$$

where v_1 and v_3 represent the two extreme points of the support of the fuzzy set and v_2 denotes the normal point where the member value reaches 1.

Recall that in Eqn. (4.1), m is the total number of rule antecedents in the rule base. Thus, and $m \geq 2$ since there is no need to assign any weight if all rules in the rule base involve just the same single antecedent attribute. The term $(m-1)$ in the first part of this formula is for local weight normalisation purpose, but it is cancelled out in the overall equation. In so doing, those n closest rules whose antecedent attributes are deemed more significant (than the rest) will be selected with priority. This is because such attributes will make less contribution (i.e., $(1 - W_j)d(A_j^p, A_j^*), j = 1, \ldots, m$) to the aggregated distance $\tilde{d}(r^p, o^*, W)$ given their relatively larger weight values.

The computation of the distance $\tilde{d}(r^p, o^*, W)$ is carried out to measure the relative closeness of the rules to the observation. Under the condition where there is no rule matching the given observation, the attribute weighted FRI is triggered. Hence, the aggregated distance is calculated as per Eqn. (4.1) between the individual elements of the observation and each corresponding rule antecedent attribute. Given this, those n rules that have resulted in the n smallest distance values are selected.

Note that the normalisation term $\frac{1}{\sqrt{\sum_{t=1}^{m}(1-W_t)^2}}$ in the above is a constant and therefore, can be omitted in the process of executing fuzzy rule interpolation. This is because selecting the closest rules only requires information on the relative distance measures.

In any potential practical application of an inference system, it is important for the underlying reasoning mechanism to be robust or stable. It is therefore, interesting to examine and establish the stability of the proposed method for weighted selection of fuzzy rules. Given a certain application, once a rule base (either predefined by experts, learned from data, or a mixture of both), an input (unmatched) observation, a distance metric (generically defined as per Eqn. (4.1)), and the thresholding number of nearest neighbours are provided, the rules which are to be selected as the nearest neighbours of the observation are in general always the same subset. There is only one extreme case where the distance measures returned are identical regarding two or more rules on the same observation such that the number of the closest neighbours including these and any other rules that are of shorter distances is greater than the given threshold. In this case, a random choice of a subset of the closest rules is made till the total number of the selected being the set threshold. Of course, the likelihood of such extreme cases taking place is very low. This ensures that in general, the proposed method is robust (see further discussion about this later in Section 5.2.2.2).

4.2 Weighted Rule Interpolation with T-FRI

In sharp contrast with conventional T-FRI techniques, the significance degrees of individual conditional attributes (captured as artificially calculated attribute weights) are herein used to compute the (interpolated) consequent given an unmatched observation. Further to the procedure for the closest rules selection as discussed above, it is naturally desirable for the resulting weights to be integrated throughout the entire interpolation process. That is, procedures for intermediate rule construction, transformation factors calculation and eventual interpolative transformation are all expected to take advantage of the weights to improve interpolation performance. Details for implementing such weighted procedures are shown below.

4.2.1 Weighted Construction of Intermediate Rule

With the attribute weighting method as introduced previously in Chapter 3, all conditional attributes can be ranked with respect to their estimated relative significance levels, reflecting their potential implication upon the derivation of the (interpolated) consequent. This allows for the development of a computational means to implement an improved version of T-FRI, where weights are integrated in all calculations during the transformation process, including the initial construction of the intermediate rule. Without unnecessarily detailing the entire construction process of the weighted intermediate rule, which is similar to that of the conventional approach (see Section 2.2.3.2), only the weighting on the consequent and the shift factor during the modified process are presented here:

$$\tilde{\hat{w}}_z^i = \sum_{j=1}^{m} W_j \hat{w}_j^i, \qquad \tilde{\delta}_z = \sum_{j=1}^{m} W_j \delta_{A_j} \tag{4.4}$$

Obviously, these will degenerate to those computed as per Eqn. (2.29) and Eqn. (2.30) in Section 2.2.3.2, when all attributes are equally regarded in terms of their significance.

4.2.2 Weighted Transformation

Given the above method for constructing the weighted intermediate rule, the scale and move factors originally provided in Eqn. (2.34) now become:

$$\tilde{s}_z = \sum_{j=1}^{m} W_j s_{A_j}, \qquad \tilde{m}_z = \sum_{j=1}^{m} W_j m_{A_j} \tag{4.5}$$

From this, if an observation that does not match any rule in the sparse rule base is presented, an interpolated fuzzy value B^* for the consequent attribute can be obtained by computing the transformation $T(\tilde{B}', \tilde{s}_z, \tilde{m}_z)$, in the exactly same way as given in Section 2.2.3.2.

4.3 Illustrative Case Study — Stage 3: Weighted T-FRI

This section continues the illustrative case study of Section 3.1.4 for weighted T-FRI, including both the process of weighted closest rule selection and that of weighted rule interpolation, when providing an observation that fails to match any rules in the rule base. To ease cross-referencing, the resulting weighted T-FRI using the weights learned by the use of information gain is hereafter referred to as IG-T-FRI, unless otherwise stated.

Recall the illustrative case where a fuzzy classification task is involved to determine which sports activity is to be undertaken (choosing from volleyball, swimming and weight lifting) given the status of four conditional attributes (i.e., temperature, outlook, humidity and wind). Suppose that the triangular membership functions adopted in this case are shown in Fig. 4.1, which are used to represent all the antecedent variables for the original data set as given in [Yuan and Shaw, 1995].

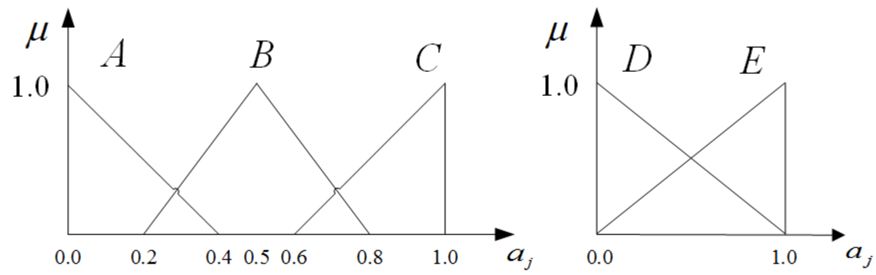


Figure 4.1: Definition of linguistic terms for domain variables.

In particular, the variables *Outlook* and *Temperature* adopt the membership functions defined on the left of the figure, where A, B, C stand for *Sunny*, *Cloudy*, *Rain*, or *Hot*, *Mild*, *Cool*, respectively in relation to the two variables. *Humidity* and *Wind* adopt the membership functions on the right, with D, E standing for *Humid*, *Normal* or *Windy*, *Not Windy*, respectively. Note that the attribute domain of each variable where the numerical values are observed has been normalised in the range of 0 to 1. The consequent variable denotes the classification outcome. Therefore, it is computationally simple to adopt the representation of singleton fuzzy sets in the description of the decision.

Suppose that the observation of Table 4.1 (involving only singleton fuzzy sets that indicate the observed numerical values) is presented, with the membership values for the observation shown in the bottom row of the table. This does not match any of the rules in the sparse rule base. Thus, no rule in the sparse rule base can be fired directly and FRI is applied to derive a conclusion. Both the weighted FRI (i.e., IG-T-FRI) and the original T-FRI are employed here for comparison.

Table 4.1: Observation in Illustrative Example

Antecedent Attribute	Temperature			Outlook			Humidity		Wind	
Observed Value	0.91			0.42			0.5		0.51	
Membership Value	*Hot*	*Mild*	*Cool*	*Sunny*	*Cloudy*	*Rain*	*Humid*	*Normal*	*Windy*	*Not Windy*
	0.0	0.0	0.775	0.0	0.733	0.0	0.5	0.5	0.49	0.51

For simplicity, the minimal number (i.e., two) of the nearest neighbouring rules are chosen for the implementation of each interpolation method. Given the rule base, the observation and the calculated attribute weights as of Table 3.5 (which are implemented with information gains), the two closest rules selected by T-FRI and those by IG-T-FRI are different, with *Rules 4* and *5* selected by T-FRI following Eqn. (2.20), and *Rules 3* and *5* by IG-T-FRI using weighted distance of Eqn. (4.1), respectively.

Using the two selected rules (*Rules 3* and *5*) via IG-T-FRI, applying the weighted T-FRI method leads to the following intermediate rule:

If *Temperature* is (0.78,0.91,1.03) and *Outlook* is (0.31,0.47,0.47) and

Humidity is (0.50,0.50,0.50) and *Wind* is (0.20,0.66,0.66),

then *Decision* is (2.49,2.49,2.49).

Differently, the intermediate rule created by the two closest rules, *Rules 4* and *5*, using T-FRI is:

If *Temperature* is (0.61,0.91,1.21) and *Outlook* is (0.42,0.42,0.42) and

Humidity is (0.50,0.50,0.50) and *Wind* is (0.01,0.51,1.01),

then *Decision* is (2.51,2.51,2.51).

Given the simplified case where observations are all singleton fuzzy sets, the above intermediate rules imply that the final interpolated result with IG-T-FRI is $\tilde{B^*} = (2.49, 2.49, 2.49)$, using the IG-guided transformation $T(\tilde{B'} = (2.49, 2.49, 2.49), \tilde{s_z} = 0, \tilde{m}_z = 0)$, and that the result with the standard T-FRI is $B^* = (2.51, 2.51, 2.51)$, using a transformation of $T(B' = (2.51, 2.51, 2.51), s_z = 0, m_z = 0)$. From this, through defuzzificaiton (to obtain a classification result), the conclusions drawn by the use of these two different methods are *Weight lifting* and playing *Volleyball*, respectively. Clearly, the outcome of applying IG-T-FRI has a better intuitive appeal given the particular observation. Indeed, recall the original rule base for this illustrative case in Section 3.1.4 that is taken from [Yuan and Shaw, 1995], the observation used for illustration actually matches *Rule 6* (i.e., the one purposefully removed to form a sparse rule base). This results in the same decision (as the ground truth) if fired as the interpolated consequent, whilst conventional T-FRI leads to an incorrect outcome.

The workflow of both the construction of the intermediate rule and the computation of the interpolative results using the original T-FRI and that using the IG-T-FRI are shown in Fig. 4.2. The left hand side of the figure illustrates the former and the right does the latter. In particular, the fuzzy sets of the antecedent variables taken by the selected closest rules, observation and the intermediate rule are shown in the first row, while the interpolated consequent values are displayed in the last. The observation of each antecedent variable and the consequent of the selected rules are both illustrated using singleton fuzzy sets for simplicity.

For T-FRI, on the left of Fig. 4.2, the fuzzy sets in dashed lines represent the variable values in *Rule 4*, while those in dash-dotted lines represent the sets in *Rule 5*. For IG-T-FRI, on the right of the figure, the fuzzy sets in dashed lines represent the variable values in *Rule 3*, and those in dash-dotted lines represent the sets in *Rule 5*. The dotted-lines represent the (fuzzy) values of the computed intermediate antecedent and consequent variables.

Note that in general and also, as in this simple case, a certain antecedent variable may not be present in both selected rules (e.g., *Temperature* and *Wind* are involved in *Rule 5* but missing in *Rule 3*). In such situations, the corresponding value of the given observation is employed to replace the missing one in the closest rules, facilitating the interpolation. This makes logical sense as the missing value of an antecedent variable in a rule indicates that any value in its domain may be matched, so the observation naturally provides the best replacement guided by the representative value. Note also that since the transformation factors in this very simple illustrative case are zero, the interpolated consequents are also shown in dotted lines in the last row of Fig. 4.2.

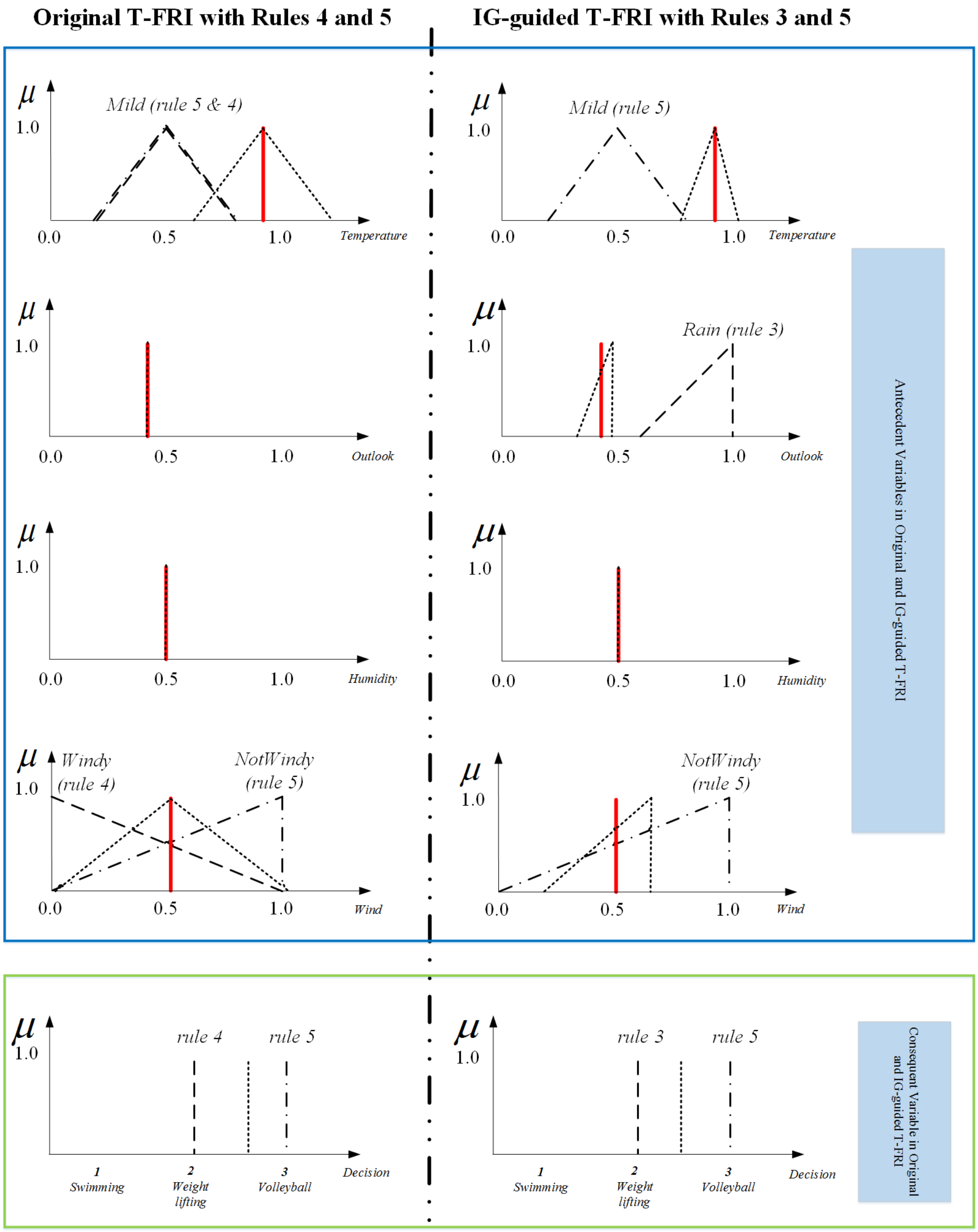


Figure 4.2: Workflow of illustrative example.

4.4 Comparison with Original Non-weighted T-FRI

As a weighted extension to the conventional T-FRI that is described in [Huang and Shen, 2006, Huang and Shen, 2008], the general rule interpolation process of this extended algorithm remains the same as its original. Note that the term of *weight* is a little over-worked herein, since it has already appeared in the conventional T-FRI, namely $w^i_j, i = 1, \dots, n, j = 1, \dots, m$ as specified in Eqn. (2.22). However, those weights are assigned for the sake of the construction of the intermediate rule, through direct comparison between the conditional attributes of a rule and the observation. This is completely different from the term of attribute weight $W_j, j \in \{1, \dots, m\}$, that is focused on here, which reveals the relative importance of each conditional attribute underpinning the original data. In particular, the weight W_j associated with a certain conditional attribute a_j is computed independent of, and fixed throughout, the interpolative process, no matter which original rule is under consideration. They are artificially calculated without acquisition of any real observations nor comparison between a given observation and any rules. Yet, in the original T-FRI, the weight w^i_j computed with respect to a certain conditional attribute is generally of a different value when a different fuzzy rule r^i is addressed.

Importantly, when all antecedent attributes are assumed to be of equal significance, namely when all weights are equal, the above modified fuzzy rule-based interpolative process degenerates to the conventional T-FRI. Mathematical proof for this is straightforward and can be easily justified. All that is required is to recall the weighting procedure regarding the individual rule antecedent attributes as described in Section 3.2. Note that the weight of each attribute has been normalised over the ranking scores derived from a given feature ranking method, which results in an identical weight for each of rule antecedent being $W_j = 1/m, j = 1, 2, \dots, m$ if all weights are assumed to be equal.

4.5 Fuzzy Rule-based Inference Supported by Weighted Interpolative Reasoning

Summing up the above developments, together with the weight learning mechanism as proposed in Chapter 3, a weighted FRI scheme can be implemented. Fig. 4.3 illustrates the workflow of such weighted T-FRI.

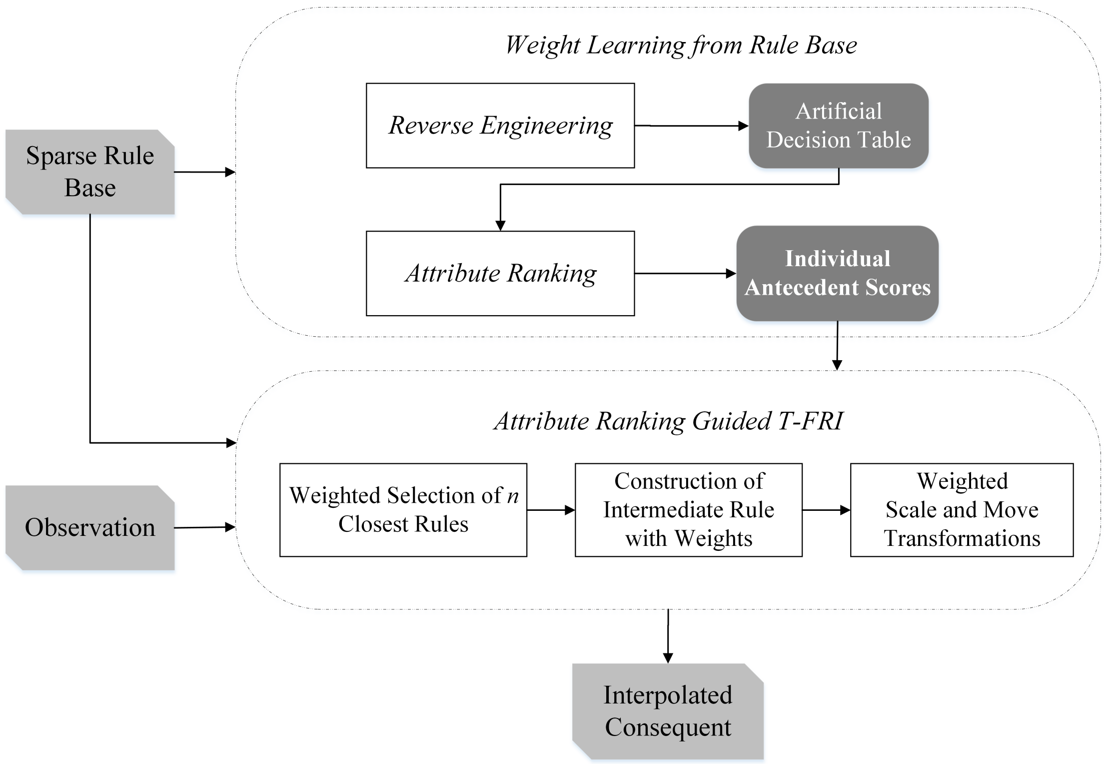


Figure 4.3: Workflow of weighted transformation-based fuzzy rule interpolation.

As indicated previously, and also clearly shown in this figure, the entire system only requires the given sparse rule base for generating the weights of rule antecedent attributes and for performing weighed fuzzy interpolative reasoning when provided with an observation. In addition, the weights which are derived from the evaluated individual antecedent scores are integrated throughout the conventional non-weighted T-FRI, covering the selection of the closest rules, the construction of an intermediate rule and the transformation process for producing the interpolated consequent. Such an integrated system works by emphasising the relative significance levels of the individual rule antecedent attributes in their use to derive more accurate interpolated consequent. The working of this weighted T-FRI has been illustrated through a case study, and further experimental verification of this will be shown later.

Traditionally, fuzzy rule-based reasoning systems infer an outcome to an unknown input or observation by firing fuzzy rules, typically using Zadeh's compositional rule of inference (CRI). Such inference works via assuming that at least one of the rules has a full or partial matching with the observation. If however, no rules have been found to match the observation, the conventional inference systems may fail to perform reasoning. Thus, extending a traditional system by emploing an FRI method

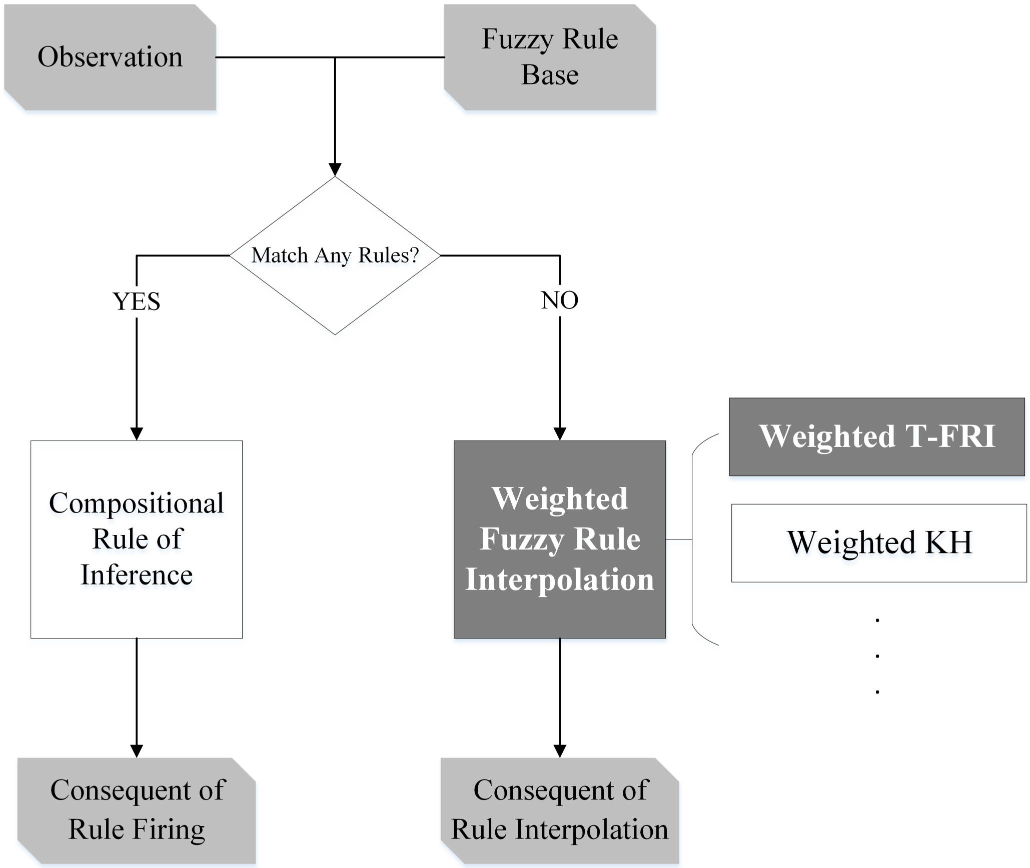


Figure 4.4: Fuzzy rule-based inference system supported by weighted FRI.

in general and the proposed attribute weighted FRI in particular will significantly reinforce the power of fuzzy rule-based systems. Fig. 4.4 shows the workflow of a generic fuzzy rule-based inference system supported by weighted rule interpolation. Note that the work so far has utilised weighted T-FRI to implement weighted interpolative reasoning mainly due to the popularity and availability of T-FRI. However, other alternative FRI approaches may also be modified with attributes weighted using the same weighting techniques. This point will be extensively addressed in Chapter 6.

Continuing the case study example employed previously, the working procedure of the above presented fuzzy rule-based inference system can be briefly illustrated also. Recall the rule base in Section 3.1.4 used for fuzzy reasoning in general and weighted FRI inference in particular for determining what sports activity to do. Suppose that here comes another observation o^* with the conditional attribute values observed as follows:

$$
\begin{array}{llll}
o^*: & \textit{Temperature} & = & 1.0, \\
 & \textit{Outlook} & = & 1.0, \\
 & \textit{Humidity} & = & 0.0, \\
 & \textit{Wind} & = & 0.2.
\end{array}
$$

Given the linguistic terms for each of the domain variables as specified in Fig. 4.1, the membership degrees of this new observation belonging to different fuzzy terms of each variable can be determined, resulting in:

$$
\begin{array}{lllll}
o^*: & \textit{Temperature} & (\textit{Hot}{=}0.0 & \textit{Mild}{=}0.0 & \textit{Cool}{=}1.0) \\
 & \textit{Outlook} & (\textit{Sunny}{=}0.0 & \textit{Cloudy}{=}0.0 & \textit{Rain}{=}1.0) \\
 & \textit{Humidity} & (\textit{Humid}{=}1.0 & \textit{Normal}{=}0.0) & \\
 & \textit{Wind} & (\textit{Windy}{=}0.8 & \textit{Not Windy}{=}0.2) &
\end{array}
$$

This means that the observation can be stated as follows:

Temperature is *Cool*, *Outlook* is *Rain*, *Humidity* is *Humid* and *Wind* is *Windy*.

When the fuzzy rule-based system illustrated in Fig. 4.4 is provided with this observation and the rule base (bar *Rule 6*) as shown in Section 3.1.4, a rule-matching check is carried out first to determine which technique (CRI or weighted FRI) to be employed next. In this example, the observation happens to match *Rule 3*, which results in the conclusion for sports activity being *Weight lifting*. If however, an observation is given in the situation as previously shown in Table 4.1, no rule has been matched, which leads to the execution of weighted FRI (more specifically, weighted T-FRI) to derive a reasonable outcome, as illustrated earlier in Section 4.3.

4.6 Summary

This chapter has dealt with the second key issue for building weighted FRI systems by integrating learned weights of rule conditionals with all important procedures of T-FRI. It has also highlighted the theoretical differences between the weighted and unweighted approaches. This resulting weighted T-FRI has been combined with classical fuzzy rule-based inference, leading to a system framework that potentially permits efficient rule firing when there is a rule matching a given observation and effective interpolation when there are no such matching rules.

Note that in the above description, no specification of which attribute ranking mechanism to use for generating the required weights is made. Indeed, the proposed technique is independent of the feature evaluation method, any of the attribute ranking methods available may be taken to assess the relative significance of individual antecedent attributes. Thus, the proposed weighted FRI offers flexibility in its implementation.

In addition, the weighted FRI system has been illustrated by the use of the same case study introduced in Chapter 3. The illustrative case is very simple but serves the purpose to explain the workflow of the weighted T-FRI. It involves only a small number of instances and a rather specific rule base. It is therefore not surprising that similar interpolated values may result by the use of either the original T-FRI or IG-T-FRI (which is one of the specific implementation of the generic framework). Nonetheless, the case demonstrates the strength of the proposed approach in that IG-T-FRI is able to produce a more intuitive outcome than the traditional unweighted T-FRI. The next chapter will systematically evaluate the work proposed so far using a range of significantly more complicated datasets over different application problems, as well as a theoretical analysis of computational complexity.

Chapter 5

Evaluation and Application of Attribute Weighted T-FRI

FUZZY rule interpolation methods have the potential in supporting reasoning in sparse fuzzy rule bases. The evaluation of them over realistic applications is essential to reveal the actual efficacy of such a system when only a sparse rule bases is available. This chapter evaluates the proposed framework for weighted fuzzy interpolative reasoning. This includes the performance evaluation of implemental systems, for learning the weights of rule antecedent attributes and more importantly, for integrating weights within the scale and move transformation-based FRI (T-FRI). Two generic application problems, classification and prediction, are taken into consideration to facilitate the evaluation, in comparison to alternative approaches.

Fuzzy interpolation techniques are desired to give prompt responses when they are implemented in time critical applications. Therefore, the complexity analysis in terms of time is a significant issue for any interpolation methods. To address this issue, the chapter will firstly analyse the computational complexity of the proposed approach from a theoretical viewpoint.

5.1 Analysis of Computational Complexity

To aid in performing computational complexity, the pseudo code of the entire weighted interpolative reasoning system is presented first. This is then followed by a systematical analysis of the time complexity.

5.1.1 Pseudo Codes of Algorithms

As reflected by Fig. 4.4 (in Section 4.5), given a fuzzy rule base R and an observation o^*, most of the conventional fuzzy rule-based systems may be able to generate a required consequent by the use of compositional rule of inference (CRI) firing the matching rule(s). If however, the rule base is sparse, where no rule matches the observation, fuzzy interpolative inference is utilised as an alternative reasoning mechanism for deriving an estimated consequent. The general framework proposed integrates both conventional CRI and a novel weighted T-FRI mechanism that is guided with the weights learned and assigned to the rule conditionals. Through this integration, it is expected to obtain more accurate inference results by exploiting the advantage of CRI for matched observations and that of the weighted FRI for unmatched ones.

The detailed methodology of weight learning from rule base and that of weighted T-FRI have been presented in Chapter 3 and Chapter 4, respectively. Alg. 5 summarises the integrated framework in pseudo code. First of all, a check is made to determine whether the observation is matched with any rule in the given rule base. If there is at least one rule being found to match the observation, the result will be obtained by firing the matched rule(s). Otherwise, the weighted T-FRI is used to make inference to estimate the consequent.

For the weighted T-FRI in particular, as illustrated in Fig. 4.3 (of Section 4.5), the weights are first learned by the use of attribute evaluation from the sparse rule base only, without requiring any observations. Then, given the rule base and the weights derived from it, the weighted T-FRI algorithm performs the required inference, through weighted search of the closest neighbouring rules of the observation and weighted interpolation with the selected closest rules. These are reflected in Lines 8 and 13 respectively, in Alg. 5, with the details for these two subroutines presented in Alg. 6 and Alg. 7 below, respectively.

Algorithm 5 Fuzzy Sparse Rule-based Inference

Input:

- Rule base $R = \{r^1, \ldots, r^N\}$, of N rules
- Observation $o^* = \{A_1^*, \ldots, A_m^*\}$, over m conditional attributes
- Cardinalities of fuzzy partitions $C = [c_1, \ldots, c_m]$, over attribute domains
- Lists of fuzzy values $F = [g_1, \ldots, g_m]$, where $g_i = \{f_1, \ldots, f_{c_i}\}$ per attribute
- Number of closest rules n

Output:

- Outcome in crisp value

for $i = 1$ to $i = N$ **do**
  Matching o^* against rule r^i;
end for
if *matched with at least one rule* **then**
  Fire matched rule(s) using CRI to obtain required consequent Z^* for o^*;
else
  if o^* is first unmatched ever **then**
    Learn weights from sparse rule base R: $W = LWFR(R, C, F)$;
    Save W;
  else
    Recall attribute weights W;
  end if
  Execute weighted FRI to compute $B^* = WeightedTFRI(R, o^*, n, W)$;
end if
Defuzzify B^* as crisp real number;
Return Crisp-valued outcome

Algorithm 6 Learning Weights from Sparse Rule Base: $W = LWFR(R, C, F)$

Input:

- Rule base $R = \{r^1, \ldots, r^N\}$, of N rules
- Cardinalities of fuzzy partitions $C = [c_1, \ldots, c_m]$, over attribute domains
- Lists of fuzzy values $F = [g_1, \ldots, g_m]$, where $g_i = \{f_1, \ldots, f_{c_i}\}$ per attribute

Output:

- Normalised attribute weights W

```
Initialise training instance pool TIP = R;
for i = 1 to i = N do
  Check if there are any missing conditionals in rule r^i;
  if no missing then
    Continue;
  else
    for each missing conditional a_k in r^i do
      Replacing r^i in TIP with c_k copies of r^i;
      Assigning a_k in each copy with one of different fuzzy values in g_k;
    end for
  end if
end for
Remove identical instances in TIP;
Calculate scores for each individual antecedent attribute using one of these:
  Score_IG = IG(TIP) or Score_Relief-F = Relief-F(TIP) or Score_LS = LS(TIP)
  or Score_LLCFS = LLCFS(TIP) or Score_RSFS = RSFS(TIP) or Score_CFS =
  CFS(TIP) or Score_IRFS = IRFS(TIP);
Calculate normalised attribute weights:
```

$$W_i = \frac{Score_*(a_i)}{\sum_{t=1,\ldots,m} Score_*(a_t)}$$

```
Return Normalised attribute weights W
```

Algorithm 7 Weighted T-FRI $B^* = WeightedTFRI(R, o^*, n, W)$

Input:

- Sparse rule base $R = \{r^1, \ldots, r^N\}$, of N rules
- Observation $o^* = \{A_1^*, \ldots, A_m^*\}$, over m conditionals
- Number of closest rules n
- Conditional weights $W = (W_1, \ldots, W_m)$

Output:

- Interpolated consequent B^*

– Closest Rules Selection:

1: **for** $i = 1$ to $i = N$ **do**

2: Calculating weighted distance $\tilde{d}(o^*, r^i, W)$ between o^* and r^i;

3: **end for**

4: Select n rules of shortest distance(s);

– Intermediate Rule (r') Construction:

5: Obtain weights $w_j^i, i = 1, \ldots, n, j = 1, \ldots, m$, as computed by original T-FRI to jth conditional attribute of ith selected rule, such that

$$w_j^i = \frac{1}{1 + d(A_j^*, A_j^i)}$$

6: Compute conditional attribute values of intermediate rule $A'_j, j = 1, 2, \ldots, m$, by linearly aggregating corresponding weighted conditional values over selected n rules using normalised weights $\hat{w}_j^i = \frac{w_j^i}{\sum_{t=1}^{n} w_j^t}$;

7: Calculate weight $\tilde{w}_z^i$ for each consequent per selected closest rule, by accumulating normalised weights contributed by $\hat{w}_j^i$, such that

$$\tilde{w}_z^i = \sum_{j=1}^{m} W_j \hat{w}_j^i$$

8: Construct fuzzy term Z' for consequent attribute of intermediate rule, by aggregating consequent values of n closest rules $z^i, i = 1, \ldots, n$, which are respectively weighted by $\tilde{w}_z^i$;

– Scale and Move Factor Calculation:

9: **for** each conditional attribute **do**

10: Obtaining scale rate s_{A_j} that modifies A'_j into $\hat{A}'_j$ such that it maintains same scale as corresponding component in o^*;

11: Obtaining move ratio m_{A_j} that modifies $\hat{A}'_j$ for it to maintain same position as corresponding component in o^*;

12: **end for**

(see next page for continuation)

(continuation of Algorithm 3)

– Scale and Move Transformation:

Calculate overall transformation factors for B' to ensure analogy, by aggregating corresponding weighted scale and move factors, such that

$$\tilde{s}_z = \sum_{j=1}^{m} W_j s_{A_j} \qquad \tilde{m}_z = \sum_{j=1}^{m} W_j m_{A_j}$$

Compute final interpolated outcome B^* by applying scale and move factors to B', such that $B^* = T(B', \tilde{s}_z, \tilde{m}_z)$;

Return B^*

5.1.2 Time Complexity Analysis

This subsection analyses the computation complexity of the proposed framework for fuzzy sparse rule based inference system which is supported by weighted T-FRI. Recall Alg. 5, 6 and 7, the time complexity of the overall approach can be estimated in the following. In particular, the two key sub-procedures, namely learning weights from sparse rule and weighted T-FRI, are analysed first, which are then collected together to derive the overall computational complexity. The notations for describing the algorithmic variables involved are the same as those specified in the **Input** statements of each algorithm.

5.1.2.1 Time Complexity of Learning Weights from Sparse Rule Base

As shown in Alg. 6, the initialisation and result return in Line 1 and Line 16 cost $O(1)$ of computation time. The **for** loop in Lines 2 - 12 repeats N times. In particular, Line 3 takes $O(m)$. Without losing generality, suppose that there are conditional attributes missing in a certain rule. Consider the worst case, where only one conditional is not missing from the rule, the **for** loop in Lines 7 - 10 repeats $m-1$ times, costing $O(c)$ for each, where $c = max\{c_1, \ldots, c_m\}$. The computation time of Line 13 involves the number of entries in the resultant training instance pool, which costs $O((size(TIP)-1)!)$. Assume that the time complexity of the method for attribute evaluation is $T(AttriEval)$, and note that the computation cost for normalised weights is $O(m)$. Thus, in total, the time complexity of Alg. 6 is

$$
\begin{aligned}
&T(LWFR) \\
&\qquad = 2O(1) + N \times [O(m) + (m-1)O(c)] + O((size(TIP)-1)!) \\
&\qquad + T(AttriEval) + O(m) \\
&\qquad = O(Nmc) + O((size(TIP)-1)!) + T(AttriEval)
\end{aligned} \tag{5.1}
$$

5.1.2.2 Time Complexity of Weighted T-FRI

In weighted T-FRI, Lines 1 - 3 of Alg. 7 cover a **for** loop which costs $N \times O(m)$ of computation time, and Line 4 takes $O(N^2)$ for sorting. Lines 5, 6 and 7 lead to a time cost of $O(mn)$ each, as they involve linear computation for every jth conditional attribute of the ith selected closest rule ($i = 1, \dots, n, j = 1, \dots, m$). Line 8 requires $O(n)$ time. Lines 9 - 12 form a **for** loop with each step in the loop (i.e., Line 10 or 11) taking a unit time of $O(1)$, and thus, the entire loop costs $O(m)$ of computation time. Line 13 takes $O(m)$ as the calculation of the transformation factors takes linear time with regard to the number of conditional attributes. Finally, the computation of the required interpolated result and returning it as shown in the last two lines take $O(1)$ time each. Note that the number of the closest rules required to perform interpolation is commonly set to $n = 2$ (or a small integer otherwise) in the existing literature (which is also experimentally justified later in this thesis). The total time complexity of weighted T-FRI is therefore, estimated to be

$$
\begin{aligned}
&T(WeightedTFRI) \\
&\qquad = N \times O(m) + O(N^2) + 3O(mn) + O(n) + 2O(m) + 2O(1) \\
&\qquad = O(N(m+N))
\end{aligned} \tag{5.2}
$$

5.1.2.3 Overall Computational Complexity

Algorithm 5 outlines the fuzzy sparse rule-based inference process, which invokes two subroutines: weights learning scheme and weighted T-FRI. Given the above analysis regarding the time complexity of these two sub-procedures, it is ready to assess the overall computational complexity of a system implementing the entire framework. The starting **for** loop in Lines 1 - 3 repeats N times (N being the number of the rules in the rule base), each of which costs $O(m)$ of computation time. The **if** statement in Line 4 takes $O(m)$ as well. Firing matched rules in Line 5 only requires a unit time of $O(1)$, otherwise, the worst case time complexity will reach the sum of $T(LWFR) + T(WeightedTFRI)$. The close up step for defuzzification and return statements costs a unit time of $O(1)$ for each. This results in the total time complexity (in the worst case):

$$
\begin{aligned}
&T_{\text{worst}} \\
&\quad = N \times O(m) + O(m) + T(LWFR) + T(WeightedTFRI) + 2O(1) \\
&\quad = O(Nmc) + O((size(TIP) - 1)!) + O(N(m+N)) + T(AttriEval)
\end{aligned}
\tag{5.3}
$$

Note that the time complexity of attribute evaluation is not detailed here as the employment of such an algorithm is independent of the FRI inference process. Naturally, an evaluation method which has less time consumption is preferred for use. As can be seen in the experimental evaluation, on practical classification and prediction applications that are to be shown next, the use of which evaluation method may not cause much difference upon the accuracy. Hence, the choice of an attribute evaluation mechanism can be made with respect to their computational simplicity.

For comparison, the time complexity of the conventional T-FRI procedure [Huang and Shen, 2006, Huang and Shen, 2008] is also checked here, which is $T(TFRI) = O(N(m+N))$. This is exactly the same as the complexity of the weighted T-FRI because the attribute weights in the weighted version are not computed within the interpolative process itself. However, regarding the entire rule-based inference system which employs just the original T-FRI for interpolative reasoning, without involving attribute weight learning, the worst total time complexity becomes: $T_{\text{worst-TFRI}} = N \times O(m) + O(m) + T(TFRI) + 2O(1) = O(mN + N^2)$, which is of course lower than that required by the weighted version and which is expected.

5.2 Evaluation with Applications to Classification and Prediction Problems

Experimental evaluation of the proposed work so far is conducted on realistic classification and prediction tasks over a range of datasets.

5.2.1 Common Experimental Set-up

Whilst different applications may involve different settings, the common set-up for both application tasks are presented first.

5.2.1.1 Fuzzy Rule Base Generation

The rule bases have been assumed to be given for the theoretical analysis of the proposed work. However, in practice, it may be difficult and even unrealistic to suggest that a rule base is readily available from domain experts. It is often required to generate rule bases in the first place for a practical application (and evaluation).

In this work, the rules used to perform both rule firing (through CRI) and rule interpolation are learned from the raw data by the use of the classical method of [Wang and Mendel, 1992], after fuzzification. The procedure of fuzzifying input variables will be explained later. Detailed procedure of this rule induction technique is summarised in Appendix A, which is employed herein forming a common ground for fair comparison. However, if preferred, more advanced rule induction mechanisms (e.g., those implemented with evolutionary algorithms) may be exploited to produce a more compact and accurate rule base.

5.2.1.2 Experimental Methodology

To minimise the potential influence of noise in judging the classification or prediction quality, experimental results are obtained by averaging the outcomes of repeated k-fold cross validation (CV) per classification or prediction dataset. In particular, the classification is conducted in 10 times 10-fold CV while the prediction is running in 10 times 5-fold CV for each dataset.

The experimental evaluation on k-fold CV is commonly used in the literature. That is, an original dataset is partitioned into k subsets of data objects, of which a single subset is retained as the testing data for the classifier, and the remaining $(k-1)$ subsets are used for training. In particular, k is often set to 5 or 10. Such CV process can be repeated for many times (say, ten). The 10 sets of results are then averaged to produce a single classifier estimation. The repeated k-fold CV is taken in terms of its advantage over random subsampling, which is that all objects are used for both training and testing in multiple times for a statistically evaluation [Qu et al., 2018].

In general, the training phase generates the rule base required for the subsequent inference (namely rule firing or rule interpolation) and the attribute weights that may be needed for the weighted interpolative reasoning, while the performance is assessed in terms of classification or prediction accuracy over the testing data. In each test, a testing sample is checked against the rules within the rule base first. If there is no rule matching the observation, FRI is applied to draw inference, using both the conventional T-FRI and the attribute weighted T-FRI to facilitate comparisons.

The weighting scheme used in classification and prediction will be specified later. Nevertheless, throughout all the experiments carried out, for feature evaluation the implementation of all the attribute weighting methods adopt the existing component tools from the Feature Selection Library (MATLAB Toolbox) [Roffo et al.,]. If desired, several parameters of these methods may be tuned in order to potentially optimise the solution for each particular problem. However, for fair comparison, the experiments conducted herein do not attempt to exhaustively tune the parameters but use the default values as set in the toolbox.

5.2.2 Classification

This section presents a systematic experimental evaluation of the proposed attribute weighted T-FRI for dealing with classification problems. It first reports on the results of performing pattern classification over ten benchmark datasets. Classification results are compared with those obtained by: (i) the state-of-the-art T-FRI; and (ii) the standard rule-based reasoning via the application of CRI, without involving rule interpolation but directly firing those (fully or partially) matched rules. Then, the robustness and effectiveness of the new approach is also empirically demonstrated by observing the following:

(i) The analysis of confusion matrices obtained for a specified case study.

(ii) The classification accuracy in relation to the number of the closest rules selected for interpolation.

(iii) The consistency and efficiency of utilising different attribute evaluation methods in implementing weighted T-FRI.

(iv) The effect of using fine-tuned membership functions in defining the fuzzy values involved in the rules.

5.2.2.1 Datasets and Particular Set-up for Classification Evaluation

A. Datasets

Ten benchmark datasets are taken from the UCI machine learning [Dheeru and Karrai, 2017] and KEEL (Knowledge Extraction based on Evolutionary Learning) dataset repositories [Alcalá-Fdez et al., 2011]. The details of these are summarised in Table 5.1. Apart from their popularity for evaluation of classification performance, these ten datasets are chosen as the attributes of different numbers are involved in different task, in order to fully assess the influence of weighting on individual attribute in fuzzy interpolative reasoning.

B. Fuzzy Values for Variables in Fuzzy Rules

As stated previously, triangular membership functions are employed here unless otherwise stated. They are used to represent the fuzzy values of the antecedent attributes. For each problem, the consequent attribute is designed to take a singleton fuzzy set (which is equivalent to a discrete crisp value), representing a certain class label. Whilst different antecedent attributes have their own underlying value domains, these domains are normalised to be within the common range of 0 to 1 and consisting of three qualitatively distinct fuzzy values, as shown in Fig. 5.1. Such a simple fuzzification is used in the main body of the experiments for simplicity as

Table 5.1: Datasets Used for Classification

Dataset	#(Attributes)	#(Classes)	#(Instances)
Iris	4	3	150
Diabetes	8	2	768
Phoneme	5	2	5404
Appendicitis	7	2	106
Magic	10	2	1902
NewThyroid	5	3	215
Banana	2	2	5300
Haberman	3	2	306
BUPA	6	2	345
Hayes-Roth	4	3	160

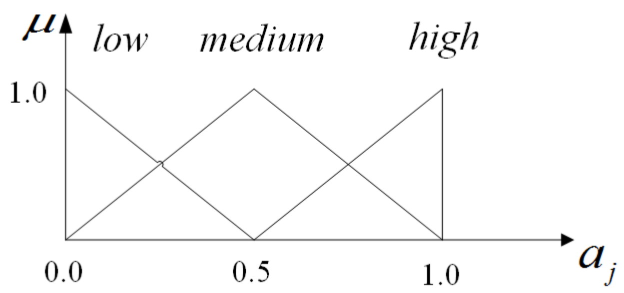


Figure 5.1: Membership functions defining antecedent attribute values for classification.

well as for fair comparison, with no optimisation of the value domain carried out. Of course, if fine-tuned membership functions are available and used, the classification performance can be expected to further improve (as to be illustrated later).

The fuzzified variables of attributes are then able to facilitate the generation of fuzzy rule bases from the data, as indicated in Section 5.2.1.1. Prior to applying the originally learned rule base to infer an outcome, an average 20% of the rules are purposefully removed randomly in order to make the resultant rule base sparser, and hence, to validate the effectiveness of rule interpolation.

C. Weighting Scheme Used for Classification

Attribute weights are derived from the use of each of the following five ranking methods (one at a time of course), Information Gain (IG), Relief-F, Laplacian Score (LS), Local Learning-based Clustering for Feature Selection (LLCFS) and Rough Set-based Feature Selection (RSFS). Such selection from the available approaches includes both types of the attribute evaluation scheme, individually or group-based owing to their availability. Given the rule base learned from the training data partitioned from the cross validation, each of the employed weighting scheme is performed using the artificial training data generated from the sparse rule base via the reverse engineering procedure.

D. Number of Closest Rules for Interpolation

The main body of this experimental study is based on the use of $n = 2$ closest rules to perform rule interpolation. However, a series of experiments are also carried out by varying the number of the closest rules selected for interpolation (see part C in the next subsection). In particular, 10 times 10-fold CV is adopted for each of the 5 different cases where the number of the closest rules selected is set to 2, 3, 4, 5, 6, respectively.

5.2.2.2 Analysis of Results

A. Classification Accuracy

Table 5.2 shows the average classification accuracies, and standard deviations (SD), which are calculated by averaging the 10 times 10-fold CV, for each of the seven compared approaches. In this table, CRI is the column showing the accuracies achievable using CRI based on the sparse rule base; Ori lists the accuracies obtained using the conventional T-FRI, with the rest naming schemes used being obvious and self-explanatory (e.g., IG stands for the accuracies achieved by the proposed approach with the antecedent attributes in the rules weighted by their corresponding information gains); and GUIDED AVERAGE presents the accuracies obtained by averaging the performances of the five attribute weighted T-FRI methods.

A comparison with the use of CRI is included herein to demonstrate the power of FRI in general and that of weighted FRI in particular in performing fuzzy reasoning, both of which significantly outperform CRI in all the problems that involve a sparse rule base. This may be expected since a fuzzy system implemented with CRI alone cannot draw any conclusion when an observation does not match any of the rules in the rule base. As already indicated, no attempt is made to optimise the fuzzification of any attribute domains. Thus, the classification rates are generally not very impressive. However, this is not the point of this experimental investigation. The point is to

Table 5.2: Average Classification Accuracies (%) with Standard Deviation over 10-times 10-fold Cross Validation

Dataset	CRI	Ori	IG	$ReliefF$	$LLCFS$	LS	$RSFS$	GUIDED AVERAGE
Iris	42.66±0.11	79.33±0.05	90.66±0.03	92.66±0.03	91.33±0.04	90.00±0.04	91.33±0.03	91.20±0.03
Diabetes	36.71±0.08	58.59±0.06	66.66±0.05	63.54±0.04	61.97±0.04	61.84±0.07	58.98±0.05	62.60±0.05
Phoneme	30.66±0.08	57.10±0.06	67.33±0.05	64.78±0.05	64.59±0.05	60.47±0.07	61.67±0.07	63.77±0.06
Appendicitis	32.08±0.10	52.00±0.16	69.72±0.15	66.91±0.19	57.72±0.17	59.45±0.14	66.72±0.17	62.11±0.16
Magic	34.76±0.09	55.84±0.05	64.35±0.03	59.77±0.05	58.94±0.06	60.89±0.06	55.84±0.04	59.96±0.04
NewThyroid	53.87±0.18	53.87±0.18	58.09±0.17	58.61±0.16	70.32±0.22	56.21±0.18	55.77±0.17	59.40±0.18
Banana	36.01±0.08	56.41±0.06	59.75±0.04	57.83±0.04	58.53±0.04	57.71±0.03	59.73±0.04	58.71±0.04
Haberman	55.44±0.10	74.26±0.11	78.83±0.08	79.15±0.11	81.11±0.09	78.50±0.09	79.15±0.12	79.35±0.10
BUPA	19.43±0.07	48.72±0.09	62.03±0.08	58.84±0.09	57.35±0.11	55.69±0.09	55.40±0.09	57.86±0.09
Hayes-Roth	36.87±0.11	46.87±0.10	60.00±0.11	60.62±0.13	54.37±0.11	56.25±0.11	58.75±0.12	58.00±0.12
Average	37.84±0.10	58.30±0.10	67.74±0.08	66.27±0.08	65.63±0.09	63.71±0.08	64.34±0.09	65.54±0.08

compare the relative performances of different approaches, provided that a common ground is ensured for fair comparison. The improvement achievable by employing learned membership functions (from training samples) will be shown later.

The use of any of the five attribute weighted methods is shown to enable the corresponding fuzzy reasoning system to outperform the system using the conventional T-FRI. This indicates that individual rule antecedent attributes do make different contributions to the classification, and that the ranking scores obtained by the feature evaluation methods taken from their original FS techniques offer positive means for discovering such differences. Interestingly, the narrow-banded SD values (those numbers following the classification accuracy) given in Table 5.2 further demonstrate that the performance of the proposed work is robust.

Examining more closely, those methods based on directly assessing individual attributes (namely, IG, Relief-F, LLCFS and LS) achieve more significant improvements, with the best average accuracy being obtained by IG-guided T-FRI (having an average improvement of 9.44% over all ten datasets than that of Ori). The remaining one, RSFS, adopts the technique of (attribute) subset selection. As shown in Section 3.2.1, ranking attributes with such a technique requires modification of the underlying FS algorithm. Nevertheless, the RSFS-based FRI has a comparable improvement over the conventional T-FRI to the average performance of the other four, again indicating the robustness of the innovative approach proposed in this work. Collectively, these results also show the generality of attribute weighted approach in that the use of a very different FS method retains the improved performance (over the underlying original T-FRI).

As also can be seen from Table 5.2, both FRI approaches (the original and the attribute weighted) significantly outperform the standard fuzzy reasoning based on CRI, and the results are more stable with a relatively lower SD values. Of course, such an obviously poorer classification accuracy obtained by the use of CRI can be expected as it fires matched or partially matched rules only while facing the problem of sparseness of the rule base. This strongly demonstrates the effectiveness of fuzzy interpolative reasoning, especially for the proposed approach owing to its further enhanced performance over the conventional T-FRI.

B. Confusion Matrices

Apart from the overall classification accuracies, it is practically interesting to investigate the statistical properties of the classification performance in terms of *true* *positive* (TP), *true negative* (TN), *false positive* (FP) and *false negative* (FN). Without overwhelming the examination while having a focused discussion, the *Haberman* dataset is taken as an example to run such an investigation. Tables 5.3-5.9 show the confusion matrices computed by the use of each of the seven compared approaches, respectively. The entries in these tables are calculated by averaging the rounded results obtained from the each 10×10 fold. Table 5.10 lists the averaged performance of the five different implementations of the attribute weighted method. Despite the fact that this dataset contains samples that are distributed in a imbalanced manner (which increases the difficulties in performing accurate classification), these tables clearly show the superior performances achieved by the proposed approach to the original T-FRI, leaving alone CRI.

Importantly, these tables reveal, both individually and collectively, that the classification accuracy achieved by the use of attribute weighted T-FRI is led by a significant reduction of false negatives and simultaneously, by a substantial increase in true positives. These results form a sharp contrast with those obtainable by the use of the original T-FRI and more remarkably, with those by CRI. This is of practical significance because for many real-world applications, not only the overall classification rates should be high, but also false negatives should be minimised while true positives are maximised. This is of particular importance for medical applications as with the situation of this dataset (which summarises the cases on the survival of patients who had undergone surgery for breast cancer – if a patient died within 5 year of the surgery then the case is regarded as positive, or if the patient survived for 5 years or longer then it is a negative case). For such problems, false negatives can be extraordinarily damaging.

Fortunately, the implementations with the proposed approach all lead to much reduced false negatives (with an averaged rate of 4.44% over the range of 3.32% to 4.88%, as compared to 8.79% returned by the conventional T-FRI and 27.04% by CRI). This is in addition to the remarkable improvements over the true positive rates (an average of 72.95% over the range of 72.32% to 74.20%, as opposite to 68.40% by the original T-FRI and a mere 50.16% by CRI).

C. Number of Closest Rules

Up till now, all experimental results reported in the existing literature regarding the use of T-FRI have been based on the use of two closest rules (i.e., $n = 2$) to

Table 5.3: Confusion Matrix of CRI

		Classified	
		Positive	Negative
Actual	Positive	50.16%	27.04%
	Negative	17.51%	5.28%

Table 5.4: Confusion Matrix of Original T-FRI

		Classified	
		Positive	Negative
Actual	Positive	68.40%	8.79%
	Negative	16.93%	5.86%

Table 5.5: Confusion Matrix of IG-T-FRI

		Classified	
		Positive	Negative
Actual	Positive	72.64%	4.56%
	Negative	16.61%	6.19%

Table 5.6: Confusion Matrix of ReliefF-T-FRI

		Classified	
		Positive	Negative
Actual	Positive	72.64%	4.88%
	Negative	15.96%	6.51%

Table 5.7: Confusion Matrix of LLCFS-T-FRI

		Classified	
		Positive	Negative
Actual	Positive	74.20%	3.32%
	Negative	15.56%	6.91%

Table 5.8: Confusion Matrix of LS-T-FRI

		Classified	
		Positive	Negative
Actual	Positive	72.32%	4.88%
	Negative	16.61%	6.18%

Table 5.9: Confusion Matrix of RSFS-T-FRI

		Classified	
		Positive	Negative
Actual	Positive	72.96%	4.56%
	Negative	16.28%	6.19%

Table 5.10: Confusion Matrix of AVGERAGE GUIDED-T-FRI

		Classified	
		Positive	Negative
Actual	Positive	72.95%	4.44%
	Negative	16.20%	6.40%

perform interpolation. The choice of using two rules is for computational simplicity. Hypotheses have been given previously in that a larger neighbourhood (i.e., more than 2 closest rules) may lead to generally more accurate interpolated outcomes. It is therefore, interesting to investigate the level of change in classification accuracy with regard to varying the number of the closest rules selected for fuzzy rule-based interpolative reasoning.

Considering the computational effort required for such an experimental investigation, only a subset of the previously listed 10 benchmark datasets (namely, BUPA, Hayes-Roth, Appendicitis and Phoneme) are randomly used to conduct this study. Table 5.11 presents the experimental results, with the summary of these plotted in Fig. 5.2. Again, the accuracies shown in in this table are calculated by averaging the results obtained over 10 times 10-fold CV.

Over the range of $n, n \in \{2, \ldots, 6\}$ that are examined, running both the conventional T-FRI and the attribute ranking-supported T-FRI always results in a substantial improvement (in terms of the classification accuracy) over the performance achievable by running CRI that works by direct rule-firing (which is shown in Table 5.2 and is irrelevant to the n). Importantly, each of the five implemented attribute-guided T-FRI methods consistently outperforms the conventional T-FRI for almost all datasets and all settings of n. The results in Table 5.11 further demonstrate the robustness of the proposed work since the standard deviation (SD) values of the classification accuracy across all n values are rather small.

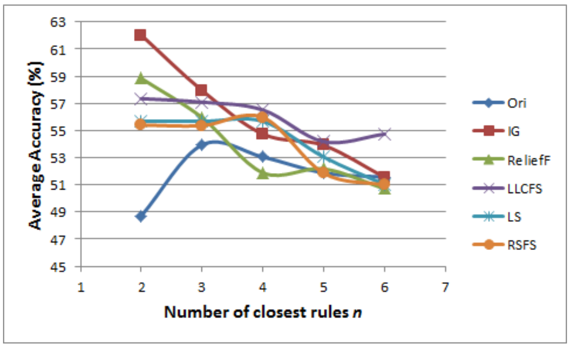
Surprisingly (and very positively in support of the present approach), as a larger n is assumed, little improvement can be gained for any of the five attribute ranking-

Table 5.11: Average Classification Accuracy (%) vs. Number of Closest Rules Used for Interpolation

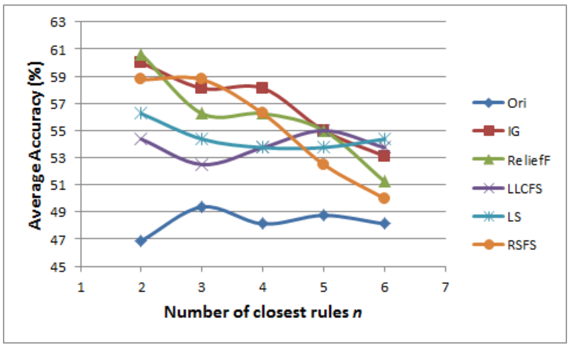
Dataset	Method	Number of Closest Rules (n) 2	3	4	5	6	SD over n
BUPA	Ori	48.72	53.95	53.05	51.88	51.57	1.98
	IG	62.03	57.95	54.74	53.91	51.56	4.05
	$ReliefF$	58.84	55.95	51.89	52.18	50.71	3.38
	$LLCFS$	57.35	57.10	56.52	54.18	54.74	1.43
	LS	55.69	55.68	55.68	53.05	51.03	2.11
	$RSFS$	55.40	55.37	55.97	51.86	50.99	2.30
Hayes-Roth	Ori	46.87	49.37	48.12	48.75	48.12	0.93
	IG	60.00	58.12	58.12	55.00	53.12	2.76
	$ReliefF$	60.62	56.25	56.25	55.00	51.25	3.35
	$LLCFS$	54.37	52.50	53.75	55.00	53.75	0.92
	LS	56.25	54.37	53.75	53.75	54.37	1.02
	$RSFS$	58.75	58.75	56.25	52.50	50.00	3.89
Appendicitis	Ori	52.00	52.18	53.00	51.09	52.09	0.68
	IG	69.72	62.09	62.18	62.18	62.18	3.38
	$ReliefF$	66.91	64.18	62.36	64.18	62.27	1.88
	$LLCFS$	57.72	56.81	56.81	55.91	55.81	0.78
	LS	59.45	55.63	54.72	53.81	53.91	2.32
	$RSFS$	66.72	63.81	62.91	60.18	60.18	2.74
Phoneme	Ori	57.10	54.16	57.54	58.91	59.19	2.01
	IG	67.33	64.93	63.45	64.63	65.08	1.40
	$ReliefF$	64.78	62.89	62.91	63.71	63.82	0.77
	$LLCFS$	64.59	61.56	60.99	60.65	61.02	1.61
	LS	60.47	61.28	60.28	61.47	62.43	0.86
	$RSFS$	61.67	61.34	61.76	61.82	60.39	0.59

supported methods. In fact, the performance may even deteriorate as n increases. The best performance is actually achieved when the number of selected closest rules is the smallest (i.e., 2). This indicates that the weighting scheme facilitates the determination of the best neighbouring rules to be taken at the earliest opportunity. This result empirically negates the hypothesis commonly made about T-FRI in that more rules used for interpolation would lead to significantly better results. It also helps avoid the use of a larger n in applications of the weighted T-FRI, thereby reducing the computational complexity that would otherwise be increased due to the requirement of searching for and running with more rules for interpolation.

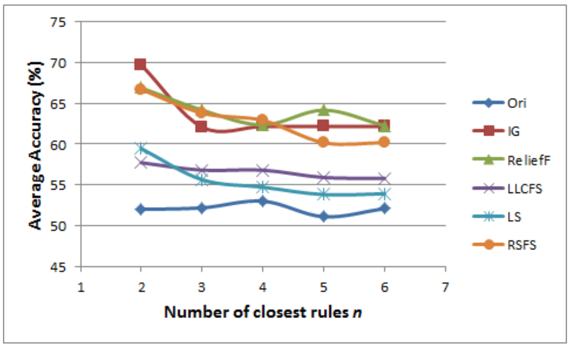
The need of just two neatest neighbouring rules helps reinforce the stability of the proposed approach. In most cases, the rules which are selected as the nearest



(a)



(b)



(c)

(d)

Figure 5.2: Accuracy variation with number of closest rules for four datasets: (a) BUPA. (b) Hayes-Roth. (c) Appendicitis. (d) Phoneme.

neighbours of the observation are identical given the availability of a certain rule base, an input (unmatched) observation, a distance metric (generically defined as per Eqn. (4.1)), and the number of nearest neighbours. There indeed exists extreme cases where a random choice of a subset of the closest rules may occur. Fortunately, the impact of such an uncertainty can be minimised as it has been empirically shown that only two nearest neighbours are required to implement the weighted FRI. That is, the employment of only two rules further minimises the already very small likelihood of having multiple rule that would return an equal distance to the observation.

D. Consistency and Efficiency of Ranking Methods

There is one exception in the above results regarding the *Phoneme* dataset where the classification accuracy achieved using LS-guided T-FRI is eventually increasing as the number of closest rules goes up, though this variation is not significant. Therefore, a further investigation has been conducted to forensically examine the ranking scores which are obtained by the use of the five different evaluation functions. The results are presented in Table 5.12.

As can be seen from this table, the first four attribute ranking methods consistently agree on that the fourth antecedent attribute plays the most significant role in deciding on the consequent, with a much higher ranking scores obtained. Three of these (IG, $ReliefF$ and $RSFS$) put the first antecedent attribute in second place, with $LLCFS$ ranking it the third. The only one method which is out of the tune is LS, which ranks the first antecedent attribute at the bottom, with a zero score signifying its relatively lack of relevancy in this rule base. This is a very different result from the great majority, implying that the LS algorithm may underperform in deriving an appropriate ranking for this particular dataset. As such, it may explain the reason that the FRI guided with LS achieves a relative poor performance when the number of the closest rules is 2 and the overall different trend of the classification accuracy while varying n in this dataset case, as shown in Fig. 5.2.

The introduction of ranking scores of antecedent attributes in support of weighted fuzzy rule interpolation may lead to additional computational overheads overall (albeit it ensuring that only the smallest number of closest rules are needed). Table 5.13 shows the corresponding average testing time recorded for classification over testing samples when the number of closest rules is increasing, together with the SD value over n. In this table, the column of Max Increase lists the maximum increase of the testing time observed while increasing the number of closest rules n.

Table 5.12: Attribute Weights and Rankings Using Different Ranking Schemes for Phoneme Dataset

Methods	Antecedent Weights					Rankings
IG	0.2852	0.0792	0.0125	0.5724	0.0507	[4 1 2 5 3]
$ReliefF$	0.1326	0.0414	0	0.7286	0.0973	[4 1 5 2 3]
$LLCFS$	0.0001	0	0	0.7416	0.2583	[4 5 1 2 3]
$RSFS$	0.0016	0.0016	0.0016	0.9938	0.0016	[4 1 2 3 5]
LS	0	0.4541	0.0988	0.1995	0.2476	[2 5 4 3 1]

Generally, there is a slight increase in time consumption when involving more closest rules in the implementation of rule interpolation for all T-FRI methods. However, whilst the attribute weighted T-FRI employs the weights in all of the key stages of interpolation (including the selection of the closest rules, the construction of the intermediate rule, the calculation of weighted transformation factors and the execution of weighted transformations), there is no significant increase in the time consumed by the weighted T-FRI as compared to that by the original T-FRI. This, together with the above observed general consistency amongst the use of different attribute ranking schemes, once again shows the efficacy of the proposed approach.

E. Use of Learned Membership Functions

As indicated previously, the classification performance in terms of accuracy is not very impressive, even though the proposed work improves it significantly over the conventional approaches. However, this is expected as the quantity space used to depict the value domains of all the attributes across all datasets is so simplistic (recall Fig. 5.1), without any optimisation (which is purposefully designed so as to enable systematic investigations over a wide range of experimental settings). Such unbiased specification of the domain values allows fair comparison to be made between different fuzzy reasoning techniques. Besides, an average of 20% of the learned rules are deliberately removed randomly, in order to have a rule base that is rather sparse. This makes the domain knowledge, represented in terms of fuzzy rules, rather incomplete, which in turn, makes the classification task a challenge for any learning classifier and hence, leads to less accurate classification. Nevertheless, it is interesting to empirically verify what if an (at least partially) optimised quantity space is utilised. In order to do this, the partition of the linguistic values of a certain rule antecedent over its domain may need to be generated by the use of data-driven methods to learn form data. The technique of clustering provides a potential solution.

Table 5.13: Average Testing Time (sec) vs. Number of Closest Rules

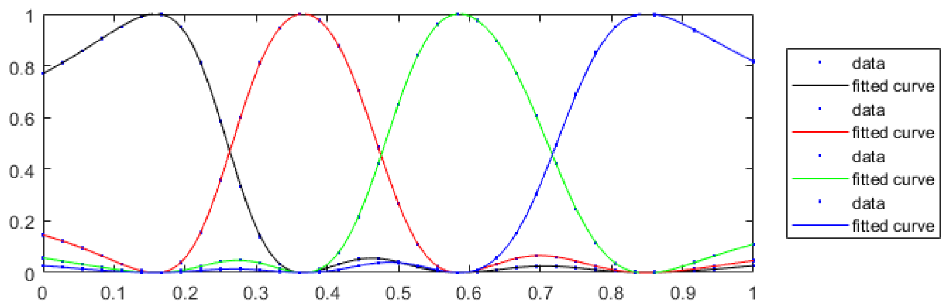
Dataset	Methods	Number of Closest Rules (n) 2	3	4	5	6	SD over n	Max Increase
BUPA	Ori	0.1584	0.1556	0.1587	0.1524	0.1624	0.0037	0.0100
	IG	0.1429	0.1445	0.1513	0.1473	0.1506	0.0036	0.0084
	$ReliefF$	0.1465	0.1448	0.1459	0.1538	0.1498	0.0036	0.0090
	$LLCFS$	0.1411	0.1473	0.1461	0.1463	0.1540	0.0046	0.0129
	LS	0.1420	0.1431	0.1446	0.1509	0.1502	0.0041	0.0089
	$RSFS$	0.1417	0.1433	0.1458	0.1464	0.1512	0.0036	0.0095
Hayes-Roth	Ori	0.0448	0.0425	0.0446	0.0439	0.0467	0.0015	0.0042
	IG	0.0386	0.0417	0.0413	0.0426	0.0425	0.0016	0.0040
	$ReliefF$	0.0394	0.0401	0.0403	0.0417	0.0408	0.0008	0.0023
	$LLCFS$	0.0404	0.0417	0.0414	0.0435	0.0426	0.0011	0.0031
	LS	0.0414	0.0415	0.0417	0.0422	0.0429	0.0006	0.0008
	$RSFS$	0.0406	0.0410	0.0420	0.0427	0.0426	0.0009	0.0021
Appendicitis	Ori	0.0334	0.0377	0.0391	0.0378	0.0394	0.0024	0.0060
	IG	0.0323	0.0350	0.0354	0.0364	0.0381	0.0021	0.0058
	$ReliefF$	0.0343	0.0347	0.0369	0.0368	0.0367	0.0012	0.0025
	$LLCFS$	0.0349	0.0352	0.0367	0.0375	0.0371	0.0011	0.0026
	LS	0.0347	0.0372	0.0368	0.0367	0.0369	0.0010	0.0022
	$RSFS$	0.0375	0.0368	0.0361	0.0369	0.0371	0.0005	0.0014
Phoneme	Ori	1.4078	1.4159	1.4448	1.4769	1.4911	0.0366	0.0833
	IG	1.3954	1.4335	1.4486	1.4728	1.4818	0.0343	0.0864
	$ReliefF$	1.3502	1.3798	1.3959	1.4303	1.4153	0.0312	0.0801
	$LLCFS$	1.3665	1.4047	1.4321	1.4328	1.4524	0.0332	0.0859
	LS	1.3440	1.3672	1.4023	1.4113	1.4256	0.0335	0.0816
	$RSFS$	1.3415	1.3900	1.3864	1.3958	1.4258	0.0302	0.0843

Fuzzy C-Means (FCM) [Bezdek et al., 1984] is one of the most widely used fuzzy clustering algorithms. It works by assigning a membership degree to each data sample corresponding to a certain cluster centre based on the relative distance between the cluster centre and that sample. The closer to the cluster centre the higher the membership degree to which the sample is deemed to belong to the corresponding cluster. Thus, the clustering outcome on a given dataset reveals the distribution of the membership functions for the underlying attributes. Owing to its popularity, FCM is herein adopted to perform fuzzification, learning the membership functions for the antecedent attributes. However, any optimisation of the membership functions is directly influenced by the dataset itself. Without overly complicating the experimental investigation, only the simple *Iris* dataset is used in this specific study (on the effect of using learned fuzzy sets).

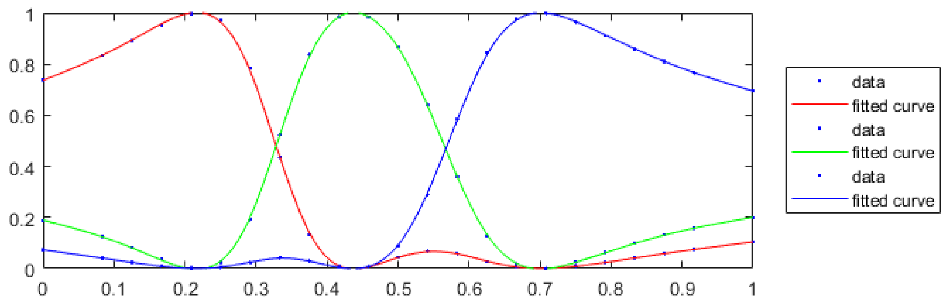
Figure 5.3 shows the membership functions generated using FCM. The optimal number of clusters for each antecedent attribute is selected by the method of [Chen and Wang, 1999], resulting in 4 clusters for the first antecedent attribute, 2 for the third and 3 for each of the remaining two.

Table 5.14 presents the classification results using the FCM-returned membership functions. For comparison, it also lists those that are obtained by the use of evenly distributed fuzzification based on the entries given in Table 5.2. As expected, a better fuzzification leads to a better classification. Individually speaking, each weighted method that uses FCM-learned membership functions beats their corresponding opponent (that employs just the simple quantity space of Fig. 5.1 for each antecedent attribute). Collectively, this leads to an averaged enhancement of 1.87% (= 93.07%−91.20%) for the FS-supported T-FRI methods. Importantly, this is on top of the already achieved substantial improvement of the FS-supported T-FRI over the conventional T-FRI and CRI-based classification methods, as also highlighted in this table.

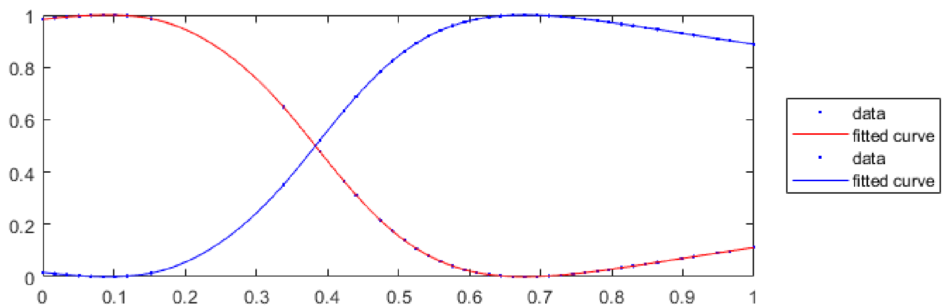
It may be recognised that the improved classification rate is still not so high as the highest possible as reported in the literature regarding this simple dataset [Riza et al., 2015], where a fully trained learning classifier is adopted with the fuzzy sets involved having been comprehensively optimised. However, it must be noticed that the present relatively high accuracy is attained with an average of 20% rules having been randomly taken out of the learned rule base. This demonstrates the great potential of the weighted FRI approach in dealing with real-world problems where typically only partial and imprecise knowledge is available.



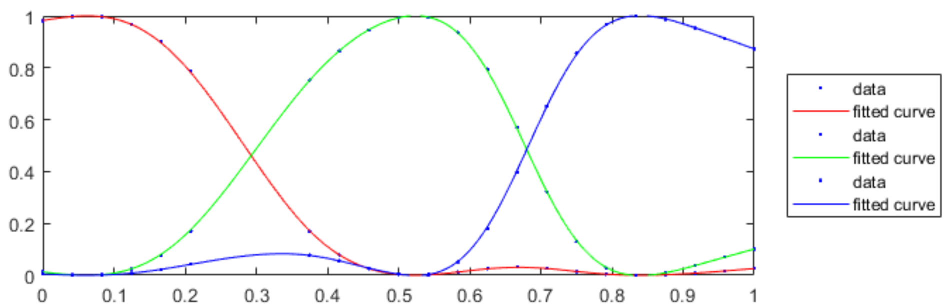
(a) Sepal Length



(b) Sepal Width



(c) Petal Length



(d) Petal Width

Figure 5.3: Membership functions learned with fuzzy c-means for Iris dataset, respectively plotted in sub-figures (a)-(d) for four attributes.

Table 5.14: Accuracies (%) vs. Specification of Membership Functions for Iris Dataset

Fuzzification Method	*CRI*	*Ori*	*IG*	*ReliefF*	*LLCFS*	*LS*	*RSFS*	AVERAGE GUIDED	Improvement over *CRI*	Improvement over *Ori*
Evenly Distributed	42.66	79.33	90.66	92.66	91.33	90.00	91.33	91.20	48.54 (114%)	11.87 (15%)
FCM-learned	68.00	88.00	95.99	93.33	92.00	90.67	93.33	93.07	25.07 (37%)	5.07 (6%)

5.2.3 Prediction

Unlike knowledge-based classification systems that use inference rules to determine categorical class labels for unknown data, prediction systems perform the forecasting of the behaviour of continuous-valued variables in a certain problem domain. Such systems enjoy a wide range of successful real-world applications, including medical case assessment [Steyerberg, 2008], object tracking and video surveillance [Ahmad et al., 2016, Sun et al., 2017], financial trend forecasting [Shin et al., 2005], civil industry simulation [Wang et al., 2016] and the generic problem of time series analysis [Box et al., 2015, Xu et al., 2018].

In this section, the proposed weighted T-FRI algorithm is applied for dealing with 12 benchmark prediction problems, including eight of which for multivariate regression and four for time series forecasting. The prediction accuracies are compared against those obtained by the conventional unweighted T-FRI method. In addition, the performance is also compared against the weighted fuzzy interpolation method as reported in [Chen and Chen, 2016], which represents the state-of-the-art of FRI involving attribute weights, across the same seven problems used in that work. Note that there were eight datasets given in [Chen and Chen, 2016] but one of which is not available for the present investigation and hence, only seven datasets are considered here.

5.2.3.1 Datasets and Particular Set-up for Prediction Evaluation

A. Datasets

The eight benchmark multivariate regression problems are taken from the popular UCI machine learning [Dheeru and Karrai, 2017] and KEEL dataset repositories [Alcalá-Fdez et al., 2011], while the four classic time series prediction problems are acquired from [Box et al., 2015, CROWDER, 1990]. These 12 datasets involve different numbers of conditional attributes and cover various real-world problem domains, including: civil engineering, energy consumption, weather forecasting, and time series prediction in industrial processes, amongst others. The properties of these datasets are summarised in Table 5.15.

Table 5.15: Datasets Used for Prediction

Dataset	#(Attribute)	#(Instance)
Abalone (KEEL)	7	4177
Concrete Compressive Strength (KEEL)	8	1030
Concrete Slump Test (UCI)	9	103
Laser (KEEL)	4	993
Plastic (KEEL)	2	1650
Daily Electricity Energy (KEEL)	6	365
Weather Izmir (KEEL)	9	1461
Auto MPG6 (KEEL)	5	392
Mackey-Glass Chaotic Time Series Prediction	4	3000
Chemical Process Concentration Readings Prediction	3	194
Chemical Process Temperature Readings Prediction	3	223
Gas Furnace Prediction	6	293

B. Fuzzy Values for Variables in Fuzzy Rules

For simplicity and consistency, the fuzzy values of all conditional attributes are again represented by triangular membership functions in this experimental investigation. The partition of each conditional attribute domain into such fuzzy values is realised by approximating what is learned by the use of Fuzzy C-Means (FCM) [Bezdek et al., 1984], owing to its popularity. The number of triangular membership functions tuned by FCM is set to 6 for each conditional attribute across all dataset, making a fair and common start point for comparison. Whilst different conditional attributes have their own underlying value domains, they are normalised to the common range of 0 to 1 before fuzzification.

In terms of rule consequent, without any prior knowledge to set an unbiased ground for comparison, the consequent learned in all prediction rules are evenly represented by isosceles triangular fuzzy sets with each having 1/5 of its domain range. The general case is illustrated with the fuzzified isosceles triangle in the middle of Fig. 5.4, where the midpoint stands for the discrete value that the predicted outcome adopts. If however, a certain prediction is so close to the boundary that one of the extreme points of the isosceles triangle locates beyond the range of the consequent domain, this extreme point is void and set to the corresponding boundary point (namely min_z or max_z) instead, as shown with both triangles on the left and right side in Fig. 5.4.

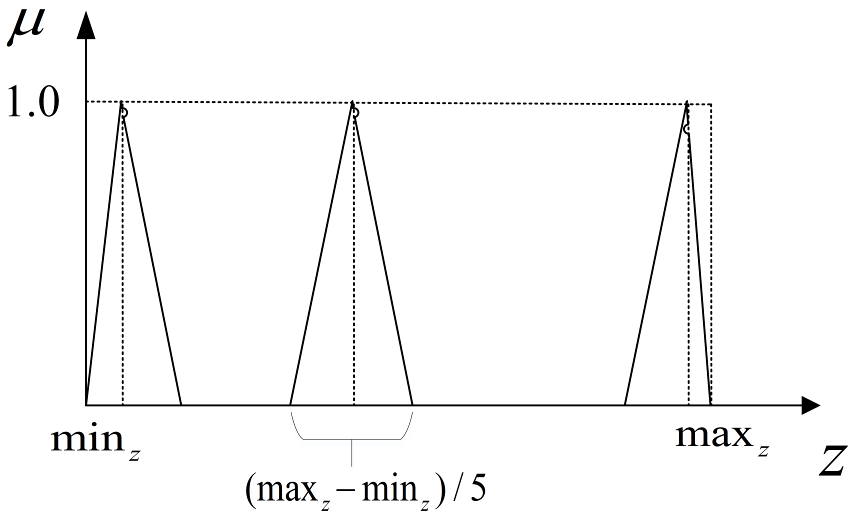


Figure 5.4: Fuzzification of consequent with isosceles triangles bounded by domain range.

C. Weighting Scheme Used for Prediction

As indicated in Section 2.4.3, the attribute evaluation functions can be broadly categorised into supervised schemes and unsupervised schemes. The unsupervised approaches offer more flexibility for prediction problems since the consequent attribute is not required during the attribute evaluation process. Therefore, two methods from this group, i.e., Local Learning-based Clustering for Feature Selection (LLCFS) and Laplacian Score (LS), are herein employed for performing the weight learning.

D. Performance metric

The prediction performance is measured using the root mean square error (RMSE) as defined by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{c}(y_i^* - y_i)^2}{c}} \tag{5.4}$$

where y_i^* and y_i represent the predicted and target outcomes of the testing samples $t_i, i = 1, 2, \ldots, c$, respectively, and c stands for the cardinality of the testing dataset. To obtain a defuzzified value as the predicted outcome, the classical defuzzification method that uses the centroid of the area under the output fuzzy set is employed. It is computed such that

$$z_0 = \frac{\int \mu_A(z) \cdot z \, dz}{\int \mu_A(z) \, dz} \tag{5.5}$$

where $\mu_A(z)$ denotes the membership degree for the variable z (of the universe of discourse) in fuzzy set A.

Generally, the smaller the RMSE values are, the more accurate the prediction is.

5.2.3.2 Analysis of Prediction Results

A. Prediction Accuracy

Table 5.16 shows the averaged prediction RMSEs directly computed using Eqn. (5.4), and the corresponding standard deviation (SD) values. In this table, the column under the heading of *Non-Weighted* lists the calculated RMSEs for the testing data obtained, by the use of CRI working together with the original unweighted T-FRI. The middle two, *LLCFS* and *LS*, list the RMSEs achieved by weighted T-FRI with the conditionals evaluated using either LLCFS or LS, respectively. Last but not least, the column of *AVG_Proposed* shows the average prediction RMSEs between the two attribute weighted T-FRI methods. From these RMSEs, it can be seen that across all datasets, the proposed approach outperforms the conventional T-FRI (that has now been strengthened with the use of CRI). This general result is not affected by the use of either of the two attribute weighting methods, as revealed by comparing the RMSEs obtained using LLCFS- and LS-weighted T-FRI.

The above results are measured on the predicted outcomes over different problem domains, showing different orders of the error scale. To facilitate a better comparison amongst different methods across all datasets, the RMSE and SD values in Table 5.16 are normalised into the range of [0,1] per dataset, with the averaged values calculated across all datasets presented in Table 5.17. A clearer comparison can now be made regarding the relative performances of the different methods investigated. Using either of the weighted T-FRI, the averaged RMSE is much smaller than that achievable by unweighted T-FRI. This indicates that introducing weights to individual

Table 5.16: Average RMSE and Standard Deviation over 10 Times 5-Fold Cross Validation

Dataset	Non-Weighted	LLCFS	LS	AVG_Proposed
Abalone	2.5124 ± 0.0714	2.4315 ± 0.0607	2.4963 ± 0.0658	2.4639 ± 0.0632
Concrete Compressive Strength	9.1938 ± 0.4948	8.6094 ± 0.5598	8.9908 ± 0.5207	8.8001 ± 0.5402
Concrete Slump Test	4.2263 ± 0.3960	3.9849 ± 0.3231	3.3304 ± 0.6013	3.6576 ± 0.4622
Laser	12.4856 ± 1.7795	12.0280 ± 1.4927	12.3016 ± 1.7496	12.1648 ± 1.6211
Plastic	2.2664 ± 0.1483	2.1878 ± 0.1188	2.1957 ± 0.1599	2.1917 ± 0.1393
Daily Electricity Energy	0.4726 ± 0.0456	0.4539 ± 0.0433	0.4621 ± 0.0356	0.4580 ± 0.0394
Weather Izmir	2.3232 ± 0.1077	2.2700 ± 0.1027	2.2873 ± 0.1126	2.2786 ± 0.1076
Auto MPG6	3.0698 ± 0.3662	2.9062 ± 0.2438	2.9275 ± 0.3268	2.9168 ± 0.2853
Mackey-Glass Chaotic Time Series Prediction	0.0432 ± 0.0047	0.0398 ± 0.0040	0.0384 ± 0.0030	0.0391 ± 0.0035
Chemical Process Concentration Readings Prediction	0.3331 ± 0.0225	0.3140 ± 0.0185	0.3169 ± 0.0213	0.3154 ± 0.0199
Chemical Process Temperature Readings Prediction	0.4221 ± 0.0329	0.3611 ± 0.0326	0.3736 ± 0.0189	0.3673 ± 0.0257
Gas Furnace Prediction	0.6527 ± 0.0484	0.6202 ± 0.0354	0.6226 ± 0.0374	0.6214 ± 0.0364

rule conditional attributes leads to more accurate prediction, and that the weights obtained by artificially learning from an original sparse rule base are effective for distinguishing the contributions of their corresponding attributes upon the prediction outcome. Moreover, the relatively lower SD values in Table 5.16 (those figures following the RMSEs), obtained by the use of weighted FRI systems across almost all datasets, further demonstrate the robustness of the proposed work. This is also verified by the results given in Table 5.17. This superior prediction performance conforms to the general results achievable by running weighted T-FRI that is tailored for classification problems, as shown in the preceding section.

Table 5.17: Comparison on RMSE and SD Averaged across Datasets

	Non-Weighted	LLCFS	LS	AVG_Proposed
RMSE	1.0000	0.0851	0.2897	0.1869
SD	0.7904	0.2780	0.5252	0.3999

Apart from the prediction error and its standard deviation, it is important to investigate whether the improvement of the attribute weighted approach over unweighted FRI is of statistical significance. Table 5.18 presents the p-values (in the range of [0,1]) returned from the statistical pairwise t-test between the attribute weighted (i.e., LLCFS- and LS-based) T-FRI and the conventional unweighted T-FRI. Given the null hypothesis that there is no significant difference between the two compared approaches, small values of p indicate doubt regarding such a hypothesis. As can be seen from this table, both LLCFS and LS weighted methods lead to rather small p-values for almost all datasets. In most cases, the test results reject the null hypothesis at a rather low significance level. Note that in this table, the asterisk sign ($*$) indicates that the improvement made by the LLCFS/LS-weighted T-FRI over unweighted T-FRI is validated at the 5% significance level (as commonly used). Also, in the situation where the Bonferroni correction [Rupert Jr et al., 2012, Shaffer, 1995] is applied for multiple significance testing, a number of comparisons are shown to have rejected the null hypothesis at a lower significance level of 2%. This implies that statistically, the attribute weighted T-FRI significantly outperforms the original unweighted version.

B. Comparison with State-of-the-art Weighted FRI

This part of experimental study compares the proposed work with the state-of-the-art weighted FRI mechanism, which is reported in [Chen and Chen, 2016] and

Table 5.18: P-value in Statistical Pairwise t-test Analysis

Dataset	LLCFS	LS
Abalone	0.0386(∗)	7.25×10^{-4}(∗)
Concrete Compressive Strength	0.0442(∗)	0.0005(∗)
Concrete Slump Test	0.0039(∗)	0.0027(∗)
Laser	0.2707	1.00×10^{-6}(∗)
Plastic	2.00×10^{-5}(∗)	0.0960
Daily Electricity Energy	2.98×10^{-9}(∗)	0.3530
Weather Izmir	5.24×10^{-11}(∗)	0.2988
Auto MPG6	0.0433(∗)	3.85×10^{-10}(∗)
Mackey-Glass Chaotic Time Series Prediction	0.0197(∗)	0.0323(∗)
Chemical Process Concentration Readings Prediction	0.0001(∗)	0.0032(∗)
Chemical Process Temperature Readings Prediction	0.0021(∗)	1.91×10^{-4}(∗)
Gas Furnace Prediction	0.0154(∗)	0.0139(∗)

is referred to as the CC method (or simply, CC taken after the names of its inventors) below. Table 5.19 and Fig. 5.5 show the results of RMSEs over seven prediction problems that have been used by CC, including both multivariate regression and time series prediction tasks. Note that different scales are used to present the results in Fig. 5.5, in an effort to reduce the impact of significant differences in the output domains over the different problems examined. For fair comparison, the settings regarding the partition of input and output attributes follow the same definition as indicated in the original work of [Chen and Chen, 2016].

Table 5.19: Comparison with CC on RMSE over 10 Times 5-Fold Cross Validation

Dataset	CC	Non-Weighted	Proposed
Abalone	2.45	2.5124	2.4639
Concrete Compressive Strength	13.44	9.1938	8.8001
Concrete Slump Test	5.91	4.2263	3.6576
Mackey-Glass Chaotic Time Series Prediction	0.0597	0.0432	0.0391
Chemical Process Concentration Readings Prediction	0.3248	0.3331	0.3154
Chemical Process Temperature Readings Prediction	0.2241	0.4221	0.3673
Gas Furnace Prediction	0.7035	0.6527	0.6214

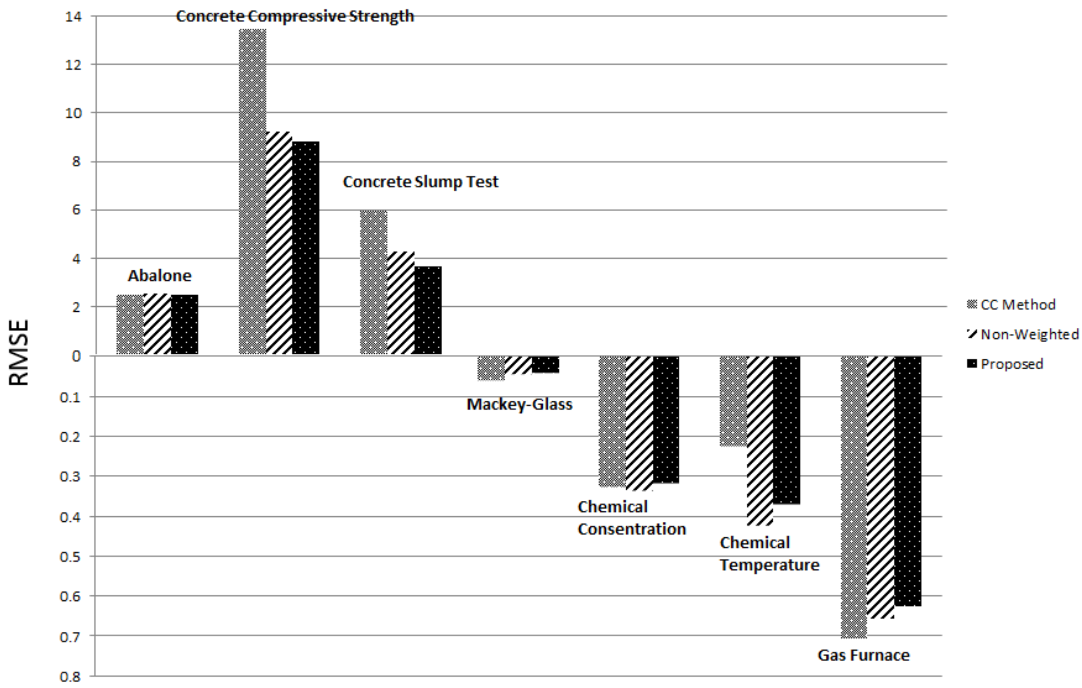


Figure 5.5: Comparison with CC on RMSE across datasets.

To minimise any potential bias against the use of a particular attribute evaluation method, the averaged performance between the two implementations of the proposed approach is shown here, in the column of *Proposed* (which is taken from Table 5.16). As empirically proven in [Chen and Chen, 2016], CC already outperforms six classical non-weighted and weighted FRI techniques in dealing with these seven prediction problems. In particular, the conventional T-FRI (which is denoted as the HS method in [Chen and Chen, 2016] without including the use of CRI) has been shown to be of less accurate performance amongst the competitors. Still, the present fuzzy sparse rule-based inference scheme, by integrating CRI (for those matched observations) and weighted T-FRI (for the unmatched ones, using the weights learned from the original sparse rule base alone), produces much more accurate results for five problems, basically ties one with CC and only underperforms with respect to CC for the dataset “Chemical Temperature”. These results can be seen in Table 5.19 and also from Fig. 5.5.

To examine the overall relative performance across all seven datasets that the compared systems have been run on, as with Part A of this subsection, normalisation on RMSEs is carried out per dataset. The resultant averaged relative RMSEs between the different approaches investigated are shown in Table 5.20. It reaffirms that the proposed approach has the smallest error in five datasets out of the seven, whilst in the other two cases it still beats the conventional unweighted T-FRI. As a whole, in comparison to CC, the averaged relative RMSE is significantly lower (0.1351 vs. 0.6472 out of a universal maximum of 1.0). The relative error reduction of 0.5121 ($= 0.6472 - 0.1351$) stands for an over 50% increase in prediction accuracy overall. Additionally, this table also shows that with an averaged RMSE reduction of 0.0876 ($= 0.6472 - 0.5596$), combining CRI and the conventional T-FRI method helps improve the performance of unweighted T-FRI to supersede that of CC, although this can be expected to certain extent given the employment of CRI. Collectively, these results positively reflect the significant potential of the proposed work.

Table 5.20: Comparison with CC on RMSE across Datasets

Dataset	CC	Non-Weighted	Proposed
Abalone	0.0000	1.0000	0.2227
Concrete Compressive Strength	1.0000	0.0848	0.0000
Concrete Slump Test	1.0000	0.2524	0.0000
Mackey-Glass Chaotic Time Series Prediction	1.0000	0.1990	0.0000
Chemical Process Concentration Readings Prediction	0.5310	1.0000	0.0000
Chemical Process Temperature Readings Prediction	0.0000	1.0000	0.7232
Gas Furnace Prediction	1.0000	0.3812	0.0000
Average	0.6472	0.5596	0.1351

5.3 Summary

In this chapter, the proposed weighted fuzzy interpolative reasoning framework has been systematically evaluated from both theoretical and practical viewpoints. To facilitate the investigation of the computational complexity of the proposed work, pseudo code of the entire system has been presented. This also helps to have a better understanding of the working process of weighted T-FRI. By applying the system to practical classification and prediction problems, this chapter has also demonstrated the very promising potential of weighted T-FRI.

Collectively, the experimental results presented have clearly shown the efficacy and robustness of the proposed approach. In particular, the weighted interpolative methods have produced results of remarkably improved classification and prediction accuracy, over both conventional T-FRI and CRI-based fuzzy reasoning techniques. This has been achieved using a very simple fuzzification mechanism. The experimental investigations have also confirmed that any feature evaluation subroutine, as a component of feature selection, may be employed to evaluate and score rule antecedent attributes, without adversely affecting the classification or prediction outcome, nor considerably increasing the computational time complexity. Experimental results have further illustrated that better performance can be obtained by fine tuning the membership functions which define the antecedent attributes within a given problem.

In addition to the aforementioned advantages over conventional T-FRI techniques, an important discovery has been achieved while performing classification evaluation. It has systematically proven that the weighted T-FRI method only requires the least number of the closest rules to carry out interpolation (with respect to a given observation that does not match any existing rule in the sparse rule base). Overall, as the most appropriate closest rules are selected in terms of the relative significance of domain attributes, better results are obtained using fewest rules possible, thereby, minimising the complexity in both rule searching and rule firing. This finding will be further evaluated by extending the current work to suit other typical FRI approaches, as to be developed next.

Chapter 6

Extensions to Attribute Weighted FRI

IN the previous chapters, a weighted interpolative reasoning scheme has been proposed, where the weights of individual antecedent attributes are learned from the given knowledge (i.e., the sparse rule base) in support of attribute ranking. Such weights are explicitly integrated with the procedures of the popular scale and move transformation-based FRI (T-FRI) [Huang and Shen, 2006]. This has led to a promising performance in tackling classification and prediction problems, as empirically shown. In particular, an important finding is that only two (i.e., the minimal number of) neighbouring rules are required for the weighted T-FRI to perform, significantly reducing the computational overheads caused by otherwise running rule interpolation with more rules.

Given this exciting empirical outcome for weighted T-FRI, it is interesting to investigate whether the discovery that "the use of least number of neighbouring rules does better" is common to other FRI methods if a similar weighting scheme is adopted. Fortunately, the weights learning mechanism as proposed in Chapter 3 is independent of the underlying FRI process, which works by exploiting the sparse rule base only.

Inspired by this observation, this chapter presents a further development that enhances two other commonly used FRI algorithms (namely, those first presented in [Kóczy and Hirota, 1993a] and [Chang et al., 2008]), by following the ideas of weighted T-FRI (which is presented in Chapter 4). The resultant weighted FRI methods are systematically evaluated via addressing ten benchmark classification problems, in comparison with their corresponding non-weighted originals. The improvement of classification accuracies is highlighted and more importantly, it is demonstrated that the best performance is achieved when the number of the nearest neighbouring rules required to perform the weighted FRI is indeed the smallest.

The rest of this chapter is structured as follows. Section 6.1 presents the modification of those two FRI methods with the use of attribute weights. Section 6.2 discusses the systematically compared experimental results. Finally, Section 6.3 provides a summary of the work reported in this chapter.

6.1 Enhancing Alternative FRI Approaches with Attribute Weighting

Two representative unweighted FRI methods (that differ from T-FRI) are considered here. These are the KH linear rule interpolation [Kóczy and Hirota, 1993a,Wong et al., 2005] and the CCL interpolation [Chang et al., 2008], which have been reviewed in Section 2.2.2.1 and 2.2.2.2, respectively. In this section, these two unweighted FRI methods are generalised by integrating the weights of rule antecedent attributes within the underlying FRI procedures. As with the presentation of the two methods previously, triangular membership functions and their associated notations for depicting the corresponding fuzzy sets, shown in Section 2.1, are adopted herein for implementing the weighted approaches, in order to maintain consistency throughout this thesis.

Note that the mechanism for learning the attribute weights $AW_j, j = 1, 2, \ldots, m$ (where m denotes the number of total rule antecedent variables appearing in the rule base), from a given sparse rule base remains exactly the same as that presented in Chapter 3 and hence, is omitted here.

6.1.1 Weighted KH Rule Interpolation

The attribute weights learned from a given sparse rule base reveal the relative significance degrees of the individual antecedent attributes, in terms of their potential in deriving the consequent given an observation. The main issue of embedding such weights within an FRI method is how to adapt the original computational mechanism of the unweighted FRI [Li et al., 2018c]. This is in order to ensure that the individually weighted attributes are aggregated in a way to better reflect their respective contributions in the interpolation process of the consequent.

As the attribute weights are learned independently of the interpolative reasoning process, all that is needed to develop a weighted version of the KH interpolation method is to modify its procedures that involve the use of α-cut distances by considering the weights accordingly. This can be carried out so that the distances are measured by taking into consideration of the relevant significance degrees of the attributes. Thus, the original unweighted KH interpolation can be extended in a straightforward manner, by computing the characteristic points of the interpolated consequent as per Eqn. (2.10) through the following weighted calculation:

$$\tilde{b_t^*} = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{\sum_{j=1}^{m} AW_j (a_{jt}^i - a_{jt}^*)^2}} b_t^i}{\sum_{i=1}^{n} \frac{1}{\sqrt{\sum_{j=1}^{m} AW_j (a_{jt}^i - a_{jt}^*)^2}}}, \quad t = 1, 2, 3 \tag{6.1}$$

Note that if the assumption of attributes having equal significance is applied, that is, $AW_j, j = 1, 2, \ldots, m$ are of the same value, the above formula degenerates to the original version, i.e., Eqn (2.10). As such, this weighted KH method is a generalised version of the original, still working as previously in the event where no weighting scheme is applicable or necessary.

6.1.2 Weighted CCL Rule Interpolation

The original CCL FRI procedure as per Section 2.2.2.2 can be generalised in a similar manner to the above. In particular, the attribute weights are integrated in the construction of the normal point b_2^* and also, in the computation of the triangular area $S_K(B^*)$ of the interpolated consequent.

In particular, the normal point b_2^* can be specified by the weighted aggregation of rule consequents of the selected neighbouring rules, where the rule weights W_i of Eqn. (2.15) are redefined by normalising the aggregated weight of each entire rule antecedent per rule. Note that in the original CCL method, the aggregation of rule weights is implemented by arithmetic average. Thus, the modified rule weight $\tilde{W_i}$ is now extended to

$$\tilde{S_K}(B^*) = \begin{cases} \left(\sum_{j=1}^{m} AW_j S_K(A_j^*)\right) \times \left(\sum_{\substack{i=1, \\ \exists j S_K(A_j^i)>0}}^{n} AW_i \times \frac{S_K(B^i)}{\sum_{j=1}^{m} AW_j S_K(A_j^i)} \right), \text{if } \exists ij,\ S_K(A_j^i) > 0 \\ \sum_{j=1}^{m} AW_j S_K(A_j^*), \quad \text{if } \forall ij,\ S_K(A_j^i) = 0 \end{cases} \tag{6.2}$$

$$\tilde{W_i} = \frac{\sum_{j=1}^{m} AW_j w_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} AW_j w_{ij}} \tag{6.3}$$

Intuitively, the average operation imposed over the rule antecedents also needs to be applied to the computation of the interpolated consequent fuzzy set. This leads to the corresponding modification of the area of the interpolated consequent fuzzy set, from Eqn. (2.14) to Eqn. (6.2). In this extension, the attribute weights $AW_j, j = 1, 2, \ldots, m$, are different from the weighting terms w_{ij} used in the original method which are still required to be computed in the same way as the original. Together, they are used to construct modified overall rule strengths. In effect, AW_j adjusts w_{ij} to better reflect the contribution of each individual antecedent attribute in relation to its significance, towards the calculation of the overall rule weight in deriving the consequent.

As with the weighted KH method, the above newly introduced rule weight $\tilde{W_i}$ and interpolated consequent area $\tilde{S_K}(B^*)$ also degenerate back to their original counterparts in the non-weighted version if all attribute weights are equal, in terms of their relative significance. Indeed, in this case, $AW_j = 1/m, \forall j \in \{1, 2, \ldots, m\}$.

6.1.3 Weighted Fuzzy Rule-based Interpolative Reasoning

Now given the weighted KH and weighted CCL interpolation methods, the fuzzy rule-based inference framework which is supported by weighted interpolative reasoning (as described in Section 4.5) can be added with new family members. The result is illustrated in Fig. 6.1, where each of all three weighted interpolation approaches provides an alternative for achieving weighted fuzzy interpolative reasoning. One and only one of them is required be to triggered when needed, of course.

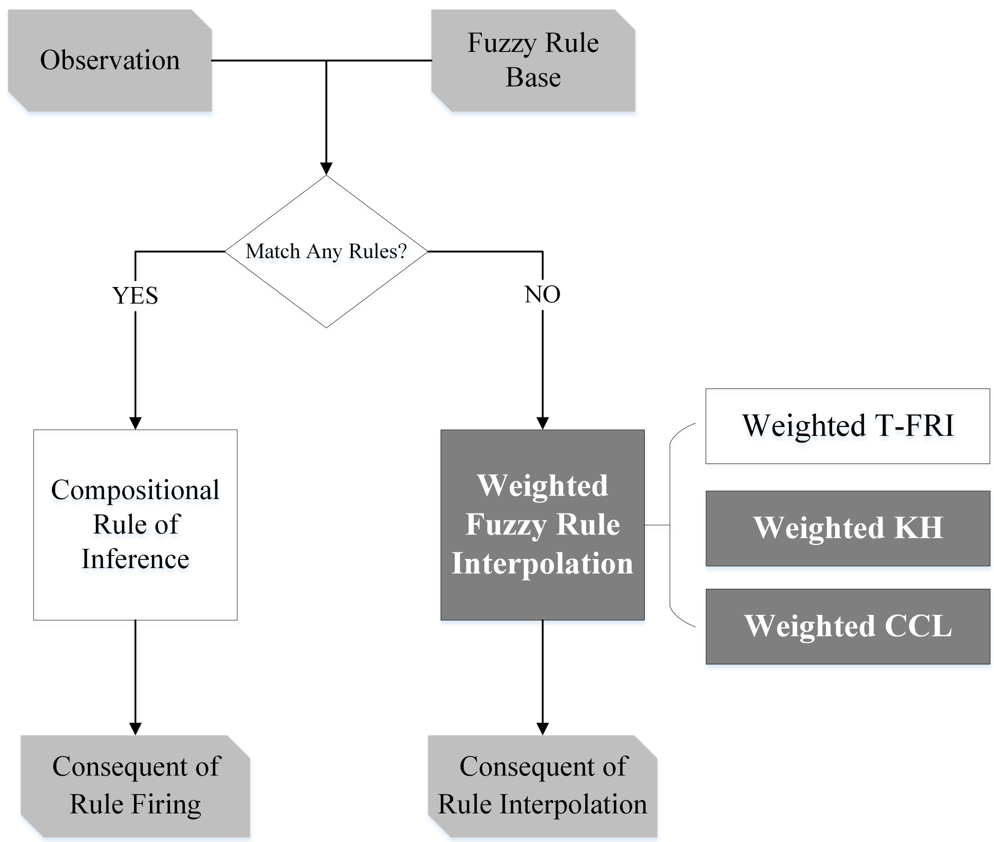


Figure 6.1: Workflow of fuzzy rule-based inference system supported by either of three weighted rule interpolation schemes.

6.2 Experimental Evaluation

This section presents a systematic experimental comparison among the proposed weighted KH, CCL and T-FRI, against their originals that do not involve individual attribute weights. The comparative investigation is performed over ten benchmark classification problems, most of which are of multiple class labels. The changes of classification accuracy with respect to the number of the nearest neighbouring rules selected for interpolation are examined, demonstrating the efficacy of weighted FRI algorithms.

6.2.1 Experimental Setup

The datasets employed for the experimentation are introduced first, followed by an outline of the experimental methodology taken.

6.2.1.1 Datasets

Ten benchmark classification datasets are taken from KEEL (Knowledge Extraction based on Evolutionary Learning) [Alcalá-Fdez et al., 2011] and UCI machine learning [Dheeru and Karrai, 2017] dataset repositories, with details summarised in Table 6.1. These ten datasets are taken for conducting the experimental evaluation of the proposed methods due to their popularity and their diversity in terms of the number of the attribute variables and that of the classes. In particular, the first five datasets are the same as those used in the experimental evaluation of Section 5.2.2 with each having a different number of attributes for binary or three-class classification, while the latter five are chosen for evaluation over many class problems.

Table 6.1: Datasets Used for Classification

Dataset	#(Attributes)	#(Classes)	#(Instances)
Diabetes	8	2	768
Phoneme	5	2	5404
Magic	10	2	1902
Haberman	3	2	306
Hayes-Roth	4	3	160
Page-blocks	10	5	5472
Ecoli	7	8	336
Red Wine Quality	11	11	1599
Wireless Indoor Localization	7	4	2000
User Knowledge Modelling	5	4	403

6.2.1.2 Experimental Methodology

As indicated previously, the proposed weighted KH, CCL and also T-FRI methods and their original versions (those given in [Wong et al., 2005], [Chang et al., 2008], and [Huang and Shen, 2006] respectively) adopt triangular membership functions to represent fuzzy values. A primitive three-valued fuzzy partition (as shown in Fig. 6.2) is employed after normalisation over all datasets, for fair comparison as well as for illustrative simplicity.

The comparative experiments are performed via 5 times 10-fold cross validation per dataset. The rule base for each problem is learned from the training data in each fold independently. The classical rule induction technique of [Wang and Mendel,

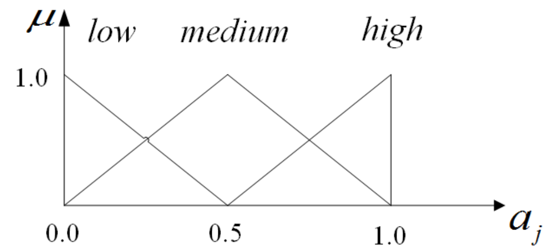


Figure 6.2: Membership functions defining values of antecedent attributes.

1992] (see Appendix A) is employed to generate an initial rule base, where 40% of the learned rules are purposefully removed randomly, resulting in a rather sparse rule base to better evaluate the performance of each FRI method. The attribute weights are then derived from the resultant sparse rule base, by the use of information gain (IG) for scoring each individual rule antecedent. Note that only IG is employed herein to compute attribute weights, because it has been shown from the performance of the weighted T-FRI with different types of weight (see evaluation results in Section 5.2) that any of the popular feature ranking methods may be utilised to perform attribute weighting without incurring much performance deviation.

For testing, as shown in Fig. 6.1, each new observation is checked against the rules in the rule base first, the consequent is calculated by aggregating the outcomes of firing the matched rules. If however, no matching is found, FRI methods are applied to derive an interpolated consequent (only one FRI method is applied at once of course, weighted or not).

Further to the comparative studies carried out between weighted FRI methods and their original unweighted ones, a series of experiments are conducted to investigate the variation of classification accuracy in relation to the number n of the nearest neighbouring rules selected for interpolation. For consistency, as with the investigation in Section 5.2.2, five different cases are compared regarding the cases where n is set to 2, 3, 4, 5, 6, respectively (whilst it makes little sense, both computationally and intuitively, to use any larger number of rules for interpolation). Also, for fair comparison, the selection scheme for the nearest neighbouring rules as described in [Huang and Shen, 2008] is employed to determine the closest rules that are required to implement the interpolation, across all six (three weighted and three original) methods compared.

6.2.2 Results and Discussion

Experimental results are presented mainly in two groups, reflecting the effectiveness and efficiency of the weighted approaches, respectively. Further results regarding classification confusion analysis and run-time performance are also reported.

6.2.2.1 Effectiveness of Weighted FRI

Table 6.2 shows the classification accuracies calculated by averaging the outcomes of 5 times 10-fold cross validation, for each of the six methods: three originals and three extended methods enhanced with the weighting scheme. The performances of weighted methods are directly compared against those of their originals, where two nearest neighbouring rules to the testing observation are selected for interpolation (unless otherwise stated). The results are presented in the column of *Weighted* and that of *Original*, respectively.

As indicated previously, a significant portion (40%) of rules are randomly removed from the original learned rule base for each classification problem, in order to thoroughly compare the performance of weighted interpolation against the unweighted. In so doing, more opportunities may be generated for those observations that find no rules to match. However, for potential practical applications, it is also desirable to investigate how much better the proposed methods do their job than their originals if more rules are available. For this purpose, Table 6.2 shows not only the comparative results obtained when the FRI methods are applied using artificially created sparse rule bases, but also the outcomes when they work with the entire learned rule bases.

As reflected in this table, by comparing the two right-most columns, not very significant improvement is gained by the weighted FRI methods if the full rule bases are employed. This can be expected as most new observations may match certain rules to fire in the first place. However, when the number of testing samples requiring interpolation becomes large, as per the situation of running a sparse rule base, each of the three weighted FRI methods significantly outperforms its corresponding unweighted method for almost all datasets.

Table 6.3 lists the (rounded) average numbers of testing samples that are unmatched by the sparse rule bases and those unmatched by the original rules. Despite

Table 6.2: Average Classification Accuracies (%) by Interpolation with Two Nearest Neighbouring Rules

Dataset	FRI	Sparser Rule Base		Full Rule Base	
		Original	Weighted	Original	Weighted
Diabetes	T-FRI	61.19	68.98	63.85	65.13
	KH	59.07	65.11	62.37	62.87
	CCL	60.92	66.83	63.91	64.61
Phoneme	T-FRI	53.89	65.85	65.51	66.18
	KH	58.07	60.79	64.22	64.41
	CCL	63.21	66.48	65.83	65.93
Magic	T-FRI	63.86	69.16	69.03	69.60
	KH	64.39	67.17	69.37	69.71
	CCL	65.71	68.43	69.51	69.74
Haberman	T-FRI	72.49	77.19	74.02	74.54
	KH	69.21	73.39	72.99	73.52
	CCL	70.91	74.77	73.93	74.46
Hayes-Roth	T-FRI	46.87	61.00	55.25	56.62
	KH	47.75	58.00	54.87	56.75
	CCL	46.37	55.75	55.50	56.75
Page-blocks	T-FRI	66.77	72.13	69.78	69.80
	KH	65.18	69.93	70.13	70.15
	CCL	66.71	72.07	69.76	69.76
Ecoli	T-FRI	59.52	65.96	62.50	65.86
	KH	59.89	64.04	62.71	65.81
	CCL	61.56	65.90	63.12	66.31
Red Wine Quality	T-FRI	52.98	57.37	52.54	53.89
	KH	52.52	53.92	52.25	53.62
	CCL	52.73	53.33	52.44	53.32
Wireless Indoor Localization	T-FRI	76.36	79.89	79.34	80.03
	KH	77.50	78.85	79.90	80.87
	CCL	75.59	77.18	78.84	79.85
User Knowledge Modeling	T-FRI	74.85	82.54	74.97	78.24
	KH	69.63	75.24	71.14	73.96
	CCL	70.53	74.22	71.52	74.18
Average	T-FRI	62.87	70.01	66.68	67.99
	KH	62.32	66.64	65.99	67.17
	CCL	63.42	67.50	66.43	67.49

the fact that there are significantly larger numbers of unmatched rules in the cases where a sparse rule base is employed, the average classification accuracies (across the ten datasets) obtained using the weighted methods beat those achievable using the full original rule bases. From the perspective of obtaining improved classification accuracy rates, this clearly demonstrates the potential of the present work.

Table 6.3: Average Number of Testing Samples for Interpolation

Dataset	Samples Requiring Interpolation in Sparser Rule Base / Total	Samples Requiring Interpolation in Full Rule Base / Total
Diabetes	58 / 77	31 / 77
Phoneme	259 / 540	52 / 540
Magic	96 / 190	45 / 190
Haberman	10 / 31	2 / 31
Hayes-Roth	9 / 16	3 / 16
Page-blocks	207 / 547	8 / 547
Ecoli	21 / 33	15 / 33
Red Wine Quality	144 / 160	99 / 160
Wireless Indoor Localization	146 / 200	78 / 200
User Knowledge Modeling	28 / 40	18 / 40

More particularly, the average improvements of the weighted T-FRI, weighted KH and weighted CCL on all ten datasets over the unweighted ones are measured to be 7.14%, 4.32%, and 4.08%, respectively. This is statistically significance as verified by pairwise t-tests, which result in low p values as listed in the third column of Table 6.4. Again, these results show that the weighted FRI methods significantly enhance the interpolative performance of the unweighted ones, and that such superior performance is attained under the condition that only two nearest neighbouring rules are employed for interpolation.

6.2.2.2 Efficiency of Weighted FRI

The previous study on weighted T-FRI (see Section 5.2.2) produced a surprising and very positive result, discovering that the use of the minimum number of nearest neighbouring rules does better for such rule interpolation. Inspired by that discovery, this part of the experimental investigation systematically looks into the effect of varying the number of neighbouring rules used for interpolation across all three weighted methods. The investigation is carried out for all aforementioned ten datasets, using five different numbers of closest rules.

Table 6.4: P-value in Statistical Pairwise t-Test

Dataset	FRI	Ori vs. Weighted ($n = 2$)	$n = 2$ vs. $n = 3$ (Weighted FRI)
Diabetes	T-FRI	8.50×10^{-6}	4.58×10^{-4}
	KH	1.44×10^{-6}	0.0254
	CCL	3.84×10^{-6}	3.16×10^{-5}
Phoneme	T-FRI	3.26×10^{-6}	1.20×10^{-4}
	KH	6.60×10^{-6}	0.1156
	CCL	1.29×10^{-4}	5.36×10^{-4}
Magic	T-FRI	4.02×10^{-5}	0.0211
	KH	1.68×10^{-5}	0.6869
	CCL	6.14×10^{-7}	0.0719
Haberman	T-FRI	1.77×10^{-5}	0.0018
	KH	1.88×10^{-4}	0.0074
	CCL	3.48×10^{-5}	9.40×10^{-4}
Hayes-Roth	T-FRI	1.66×10^{-5}	0.0300
	KH	2.98×10^{-7}	0.0155
	CCL	2.91×10^{-5}	0.1575
Page-blocks	T-FRI	1.62×10^{-5}	1.29×10^{-4}
	KH	8.26×10^{-5}	6.82×10^{-5}
	CCL	2.50×10^{-5}	7.24×10^{-4}
Ecoli	T-FRI	1.84×10^{-6}	5.96×10^{-5}
	KH	2.80×10^{-4}	6.22×10^{-4}
	CCL	5.24×10^{-5}	1.52×10^{-4}
Red Wine Quality	T-FRI	2.03×10^{-6}	0.0025
	KH	0.0559	0.0152
	CCL	0.0423	0.2836
Wireless Indoor Localization	T-FRI	1.79×10^{-5}	0.0019
	KH	9.55×10^{-4}	0.0089
	CCL	0.0020	0.1940
User Knowledge Modeling	T-FRI	1.50×10^{-5}	2.26×10^{-4}
	KH	0.0026	0.0075
	CCL	0.0030	8.13×10^{-4}

Note that attribute weights can also be exploited to help modify the selection procedure for the nearest neighbouring rules (see Section 4.1 for details). Thus, in order to thoroughly examine the implication of the weighting scheme upon both the procedure for closest rules selection and that for rule interpolation, the experiments on classification results are herein purposefully designed to cover the following all four cases, for each particular FRI approach (be it T-FRI, KH or CCL): unweighted selection with unweighted interpolation, unweighted selection with weighted interpolation, weighted selection with unweighted interpolation, and weighted selection with weighted interpolation. These are denoted as $S_{\bar{w}}I_{\bar{w}}$, $S_{\bar{w}}I_w$, $S_wI_{\bar{w}}$ and S_wI_w respectively. Of course, if the number of the neighbouring rules is set to two, then the first and the last become exactly the same as those denoted by *Original* and *Weighted* as previously given in Table 6.2, running on a sparse rule base.

Tables 6.5 - 6.9 (with Tables 6.6 - 6.9 being the continuations of Table 6.5 due to the limit of the physical space) present the results of this set of experiments, with the examined range of n set to $\{2, 3, \ldots, 6\}$. This is partly to facilitate direct comparison with the state-of-the-art results provided in Section 5.2.2, and partly to reflect the practical consideration where using more than six closest rules to perform interpolation is of little intuitive appeal, both in terms of computational complexity and of classification result interpretability. Over this entire range, the accuracies obtained by the use of weighted interpolation generally outperform those by the unweighted for all three FRI approaches. That is in most cases, the results achieved by $S_{\bar{w}}I_w$ are improved over $S_{\bar{w}}I_{\bar{w}}$, while S_wI_w does better than $S_wI_{\bar{w}}$. These improvements further demonstrate the effectiveness of the weighted FRI methods proposed here.

Figure 6.3 plots the changing trend of classification accuracy in relation to the number of the nearest neighbouring rules used. As n goes up from the minimum (i.e., $n = 2$), the accuracies drop, sometimes sharply, for all three weighted FRI methods with the weighted interpolation supported by weighted rule selection (i.e., S_wI_w). This behaviour of weighted FRI for the *Magic* and *Red Wine Quality* datasets is slightly less obvious, but increasing n does not help to improve the classification performance either.

The observation that the results of S_wI_w with any FRI approach when $n = 2$ beat those when $n = 3$ is further validated by pairwise t-test in Table 6.4, with p values shown in the fourth column of this table. These experimental results indicate that the reduction of classification accuracies when the number of the nearest neighbouring rules is increased from 2 to 3 is statistically significant for almost all FRI methods across all datasets.

Examining the results of Tables 6.5-6.9 more closely, as highlighted in bold for each of the ten datasets, the best performance of each FRI across the four implementations (namely, $S_{\bar{w}}I_{\bar{w}}$, $S_{\bar{w}}I_w$, $S_wI_{\bar{w}}$, and S_wI_w) over the entire range of n studied, is

Table 6.5: Average Classification Accuracies (%) vs. Number of Nearest Neighbouring Rules Used for Different FRI

Dataset	FRI	n	$S_{\bar{w}}I_{\bar{w}}$	$S_{\bar{w}}I_w$	$S_wI_{\bar{w}}$	S_wI_w
Diabetes	T-FRI	2	61.19	66.50	63.69	**68.98**
		3	63.64	63.77	62.44	63.09
		4	65.15	67.13	63.87	66.00
		5	64.68	65.41	63.12	65.02
		6	65.12	66.76	64.21	65.90
	KH	2	59.07	63.81	60.21	65.11
		3	63.52	63.49	61.95	62.03
		4	63.00	64.67	61.77	63.08
		5	64.46	64.51	62.50	62.87
		6	64.28	**65.37**	62.45	63.49
	CCL	2	60.92	64.90	62.71	**66.83**
		3	62.24	62.63	59.87	60.31
		4	63.65	64.74	61.80	63.15
		5	63.39	63.65	61.72	61.56
		6	64.01	65.00	61.93	63.28
Phoneme	T-FRI	2	53.89	56.43	62.15	**65.85**
		3	55.09	55.22	58.57	62.37
		4	54.78	57.08	61.09	64.51
		5	55.59	56.74	57.75	64.52
		6	56.11	58.51	60.86	65.33
	KH	2	58.07	59.53	59.21	**60.79**
		3	59.50	59.50	59.38	59.38
		4	59.58	59.78	59.47	59.42
		5	59.34	59.40	59.35	59.39
		6	60.03	60.12	59.91	60.19
	CCL	2	63.21	64.31	65.68	**66.48**
		3	58.84	60.82	60.59	62.33
		4	63.57	64.05	64.44	64.83
		5	60.22	63.27	60.81	63.79
		6	63.90	65.06	64.49	65.32

generally achieved using S_wI_w with $n = 2$. However, for the conventional $S_{\bar{w}}I_{\bar{w}}$ FRI methods where no weighting scheme is employed, the accuracy increases with n. This forms a sharp contract between the weighted and unweighted approaches, and demonstrating the efficacy of the proposed work.

There are exceptional cases to observe. Particularly, the results show that the $S_{\bar{w}}I_w$ FRI method can do better than the rest if a large number (e.g., $n = 5$ or $n = 6$) of rules

Table 6.6: Average Classification Accuracies (%) vs. Number of Nearest Neighbouring Rules Used for Different FRI (Continued)

Dataset	FRI	n	$S_{\bar{w}}I_{\bar{w}}$	$S_{\bar{w}}I_w$	$S_wI_{\bar{w}}$	S_wI_w
Magic	T-FRI	2	63.86	67.86	65.46	69.16
		3	67.79	67.97	67.53	67.95
		4	67.37	68.90	67.30	68.95
		5	68.91	69.02	68.28	68.90
		6	68.78	**69.61**	68.23	69.45
	KH	2	64.39	66.29	65.01	67.17
		3	66.79	66.78	67.00	67.05
		4	66.95	66.91	66.89	67.00
		5	67.45	**67.47**	67.31	67.44
		6	67.45	67.43	67.18	67.43
	CCL	2	65.71	67.62	66.48	68.43
		3	67.58	67.78	67.57	67.81
		4	67.41	68.03	67.16	67.79
		5	**68.48**	68.29	68.17	67.91
		6	67.83	68.22	67.31	67.73
Haberman	T-FRI	2	72.49	74.44	75.36	**77.19**
		3	72.94	73.39	73.53	74.56
		4	74.38	74.77	74.19	74.25
		5	74.32	74.83	74.12	74.71
		6	74.25	74.58	74.58	74.44
	KH	2	69.21	70.26	72.28	**73.39**
		3	72.29	71.69	72.88	71.30
		4	72.21	72.35	72.35	72.42
		5	72.16	71.64	72.55	71.36
		6	72.69	71.90	72.29	71.95
	CCL	2	70.91	72.36	73.13	**74.77**
		3	71.77	72.62	72.15	72.68
		4	72.48	72.88	71.83	72.29
		5	71.97	72.30	72.09	72.29
		6	72.09	71.97	71.90	71.90

are used. Such situations occur mostly when the KH weighted interpolation method is employed with rules taken by unweighted selection. Nonetheless, the interpolation procedure is still weighted in these cases; this again shows the effectiveness of weighting upon rule antecedent attributes. Besides, there is little win of $S_{\bar{w}}I_w$ over S_wI_w. Yet, such minor win is obtained at the expense of much more computational overheads as more rules are involved in the interpolation procedure, as illustrated below. This finding is of great importance in practical application of FRI since it

Table 6.7: Average Classification Accuracies (%) vs. Number of Nearest Neighbouring Rules Used for Different FRI (Continued)

Dataset	FRI	n	$S_{\bar{w}}I_{\bar{w}}$	$S_{\bar{w}}I_w$	$S_wI_{\bar{w}}$	S_wI_w
Hayes-Roth	T-FRI	2	46.87	47.00	60.37	**61.00**
		3	48.75	48.75	55.75	56.37
		4	50.50	50.50	53.99	53.37
		5	52.50	52.37	55.75	55.87
		6	51.87	52.37	54.75	54.87
	KH	2	47.75	48.50	57.12	**58.00**
		3	49.75	49.62	52.37	51.62
		4	51.37	52.00	51.87	52.12
		5	51.00	51.37	51.87	52.50
		6	51.12	51.12	51.50	51.50
	CCL	2	46.37	47.12	55.25	**55.75**
		3	48.25	49.00	51.87	53.25
		4	50.62	50.37	53.00	52.24
		5	52.37	51.62	53.00	52.37
		6	50.87	51.12	51.50	51.87
Page-blocks	T-FRI	2	66.77	66.76	72.12	**72.13**
		3	60.54	59.26	61.25	60.23
		4	65.45	65.37	64.99	64.91
		5	66.81	65.66	66.57	65.49
		6	64.97	64.28	64.39	63.85
	KH	2	65.18	65.12	69.86	**69.93**
		3	57.83	57.83	58.22	58.23
		4	59.39	58.02	59.11	57.82
		5	60.30	60.14	60.11	59.93
		6	59.44	58.83	59.00	58.58
	CCL	2	66.71	66.68	72.03	72.07
		3	**74.67**	63.77	74.63	64.63
		4	71.80	66.05	72.27	65.92
		5	68.08	65.86	68.31	65.86
		6	70.02	65.06	69.12	64.71

empirically confirms that weighted FRI methods only require two (i.e., the least number of) nearest neighbouring rules to perform rule interpolation, significantly enhancing the overall algorithm efficiency.

Table 6.8: Average Classification Accuracies (%) vs. Number of Nearest Neighbouring Rules Used for Different FRI (Continued)

Dataset	FRI	n	$S_{\bar{w}}I_{\bar{w}}$	$S_{\bar{w}}I_w$	$S_wI_{\bar{w}}$	S_wI_w
Ecoli	T-FRI	2	59.52	61.38	64.34	**65.96**
		3	56.56	56.50	55.77	55.71
		4	52.87	52.51	52.15	51.85
		5	49.04	48.93	48.99	49.11
		6	47.38	47.85	46.06	46.24
	KH	2	59.89	59.82	63.92	**64.04**
		3	56.50	56.50	55.77	55.77
		4	53.40	53.11	52.09	51.97
		5	49.52	49.53	49.17	49.12
		6	48.44	48.50	47.07	47.08
	CCL	2	61.56	61.68	65.60	**65.90**
		3	56.50	56.50	55.71	55.77
		4	52.87	52.81	52.34	52.46
		5	49.10	49.22	49.17	49.29
		6	47.37	47.37	46.06	46.07
Red Wine Quality	T-FRI	2	52.98	55.64	54.55	**57.37**
		3	53.29	53.28	53.17	53.14
		4	53.23	54.24	53.27	54.11
		5	53.33	53.39	52.69	52.85
		6	53.68	54.19	53.10	53.42
	KH	2	52.52	52.83	53.22	53.92
		3	53.26	53.23	53.14	53.07
		4	53.43	53.84	53.48	53.78
		5	53.29	53.29	52.68	52.62
		6	53.81	**54.13**	53.29	53.28
	CCL	2	52.73	52.74	53.24	53.33
		3	52.62	53.14	52.14	52.89
		4	53.56	53.14	53.37	52.99
		5	53.84	**54.54**	53.29	53.68
		6	53.87	53.87	52.95	52.55

Table 6.9: Average Classification Accuracies (%) vs. Number of Nearest Neighbouring Rules Used for Different FRI (Continued)

Dataset	FRI	n	$S_{\bar{w}}I_{\bar{w}}$	$S_{\bar{w}}I_w$	$S_wI_{\bar{w}}$	S_wI_w
Wireless Indoor Localization	T-FRI	2	76.36	78.32	77.94	**79.89**
		3	77.22	77.22	76.93	76.93
		4	75.12	75.45	74.56	74.91
		5	74.41	74.50	73.92	74.02
		6	76.47	77.00	75.92	76.44
	KH	2	77.50	78.02	78.37	78.85
		3	77.22	77.22	76.93	76.94
		4	75.50	75.89	75.03	75.25
		5	74.52	75.10	74.06	74.79
		6	79.33	**79.62**	79.06	79.06
	CCL	2	75.59	76.24	76.12	**77.18**
		3	76.82	76.95	76.53	76.89
		4	75.37	76.12	74.66	75.58
		5	76.36	77.12	75.89	76.82
		6	76.11	76.33	75.47	75.76
User Knowledge Modeling	T-FRI	2	74.85	78.91	77.03	**82.54**
		3	74.44	75.54	63.84	69.55
		4	76.18	78.91	66.17	69.15
		5	76.29	79.16	61.55	68.50
		6	77.92	79.95	61.10	66.85
	KH	2	69.63	71.51	72.68	75.24
		3	74.69	74.69	64.89	68.21
		4	74.05	74.39	65.73	69.05
		5	76.09	**76.68**	62.94	68.01
		6	75.78	76.02	62.04	66.90
	CCL	2	70.53	69.73	70.64	74.22
		3	73.64	71.91	63.84	66.02
		4	73.99	72.05	63.19	66.65
		5	74.80	74.34	60.60	66.06
		6	**75.64**	74.54	59.50	64.42

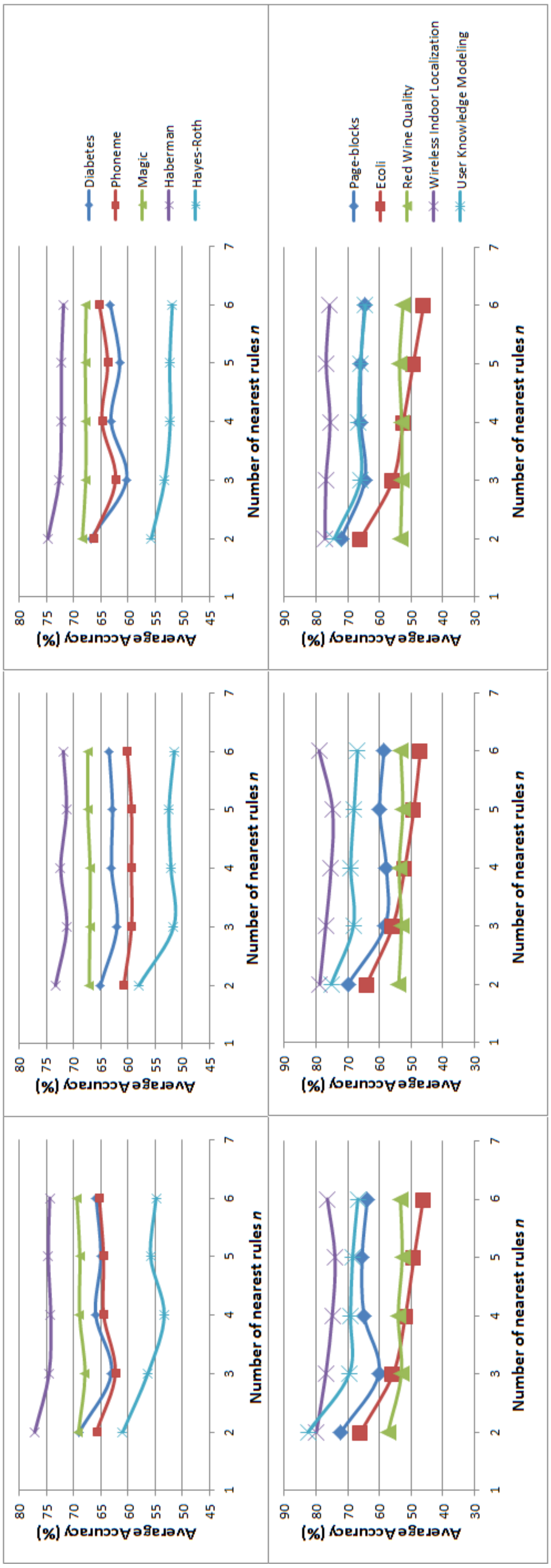


Figure 6.3: Accuracy variation in response to number of neighbouring rules.

6.2.2.3 Further Analysis

A. Confusion Matrix

The analysis of confusion matrices has also been conducted for each of the three weighted FRI methods regarding the use of two or three nearest neighbouring rules. To save space, Tables 6.10-6.12 present the outcomes for the *Diabetes* dataset as an example case study, since the general trends for the others are similar. The comparison in each of these tables helps explain why the overall classification accuracy may dramatically decrease as n increases from 2 to 3. As reflected by these results, the adverse variation of the overall accuracy when $n = 3$ appears to be caused by the significant increase of false positives and the considerable reduction of true negatives. Of course, such situations must be minimised in any realistic application, especially for instance in medical diagnosis as is indeed the case concerning this dataset. Both increase in false positives and reduction in true negatives will usually cause undue anxiety of the patient, and in worse scenarios, may even cause missing the correct diagnosis of other disease(s) that the patient may be suffering from the given symptoms.

B. Run Time

Results so far have demonstrated that weighted FRI methods (that involve additional computation in both rule selection and rule interpolation procedures) generally outperform their originals. However, a question may be raised as to how much extra computation effort is required to attain such improved performance, despite the recognition that learning the weights themselves is an offline task. This final experimental study therefore, addresses this natural concern regarding the run time performance of the weighted methods.

Table 6.13 lists the average testing time recorded for all three weighted FRI methods (i.e., in the form of $S_w I_w$) and their originals (namely, $S_{\bar{w}} I_{\bar{w}}$), when dealing with the final five problems given in Table 5.1. The tests are carried out in relation to the increase of the number of the nearest neighbouring rules employed. Note that these five cases are selected because they each involve more classes and hence, are more difficult to classify (whilst saving the space otherwise required to present similar results for the other five). As expected, there is indeed an increase in time consumption when exploiting more nearest neighbouring rules for all FRI methods

Table 6.10: Confusion Matrix of Weighted T-FRI with $n = 2$ and $n = 3$ Nearest Neighbouring Rules

		Classified ($n = 2$)		Classified ($n = 3$)	
		Positive	Negative	Positive	Negative
Actual	Positive	23.12%	11.76%	21.09%	13.79%
	Negative	19.26%	45.82%	24.34%	40.75%

Table 6.11: Confusion Matrix of Weighted KH with $n = 2$ and $n = 3$ Nearest Neighbouring Rules

		Classified ($n = 2$)		Classified ($n = 3$)	
		Positive	Negative	Positive	Negative
Actual	Positive	20.70%	14.19%	19.99%	14.89%
	Negative	20.74%	44.34%	23.57%	41.51%

Table 6.12: Confusion Matrix of Weighted CCL with $n = 2$ and $n = 3$ Nearest Neighbouring Rules

		Classified ($n = 2$)		Classified ($n = 3$)	
		Positive	Negative	Positive	Negative
Actual	Positive	18.14%	16.73%	18.92%	15.95%
	Negative	16.49%	48.60%	23.65%	41.44%

(weighted or not). The use of fewer rules will thus be more efficient. However, as can be seen from this table, there is no significant increase in the time cost by a weighted FRI as compared to that by its original where no weights are involved, while using the same number of rules for interpolation. This once again demonstrates the efficacy of the proposed weighted FRI techniques and supports the outcome that least neighbouring rules do better with attribute weighted FRI.

6.3 Summary

This chapter has further developed the work of Chapter 4 on weighted fuzzy interpolative reasoning, by extending the weighted transformation-based FRI to two other classical FRI methods, namely the KH [Kóczy and Hirota, 1993a] and CCL [Chang

Table 6.13: Average Testing Time (sec) vs. Number of Nearest Neighbouring Rules

Dataset	Methods		Number of Closest Rules (n) 2	3	4	5	6
Page-blocks	T-FRI	$S_{\bar{w}}I_{\bar{w}}$	0.1881	0.1930	0.1953	0.2013	0.2049
		S_wI_w	0.1876	0.1924	0.1961	0.2045	0.2063
	KH	$S_{\bar{w}}I_{\bar{w}}$	0.1794	0.1813	0.1915	0.1940	0.1971
		S_wI_w	0.1829	0.1839	0.1910	0.1941	0.1972
	CCL	$S_{\bar{w}}I_{\bar{w}}$	0.1789	0.1784	0.1833	0.1881	0.1887
		S_wI_w	0.1814	0.1813	0.1854	0.1922	0.1925
Ecoli	T-FRI	$S_{\bar{w}}I_{\bar{w}}$	0.0165	0.0203	0.0192	0.0207	0.0206
		S_wI_w	0.0162	0.0199	0.0192	0.0217	0.0207
	KH	$S_{\bar{w}}I_{\bar{w}}$	0.0166	0.0200	0.0175	0.0201	0.0198
		S_wI_w	0.0164	0.0197	0.0177	0.0199	0.0184
	CCL	$S_{\bar{w}}I_{\bar{w}}$	0.0181	0.0182	0.0188	0.0209	0.0199
		S_wI_w	0.0184	0.0198	0.0184	0.0213	0.0201
Red Wine Quality	T-FRI	$S_{\bar{w}}I_{\bar{w}}$	0.1673	0.1735	0.1733	0.1782	0.1819
		S_wI_w	0.1649	0.1722	0.1729	0.1773	0.1799
	KH	$S_{\bar{w}}I_{\bar{w}}$	0.1687	0.1695	0.1736	0.1754	0.1784
		S_wI_w	0.1692	0.1713	0.1753	0.1747	0.1777
	CCL	$S_{\bar{w}}I_{\bar{w}}$	0.1609	0.1612	0.1632	0.1647	0.1684
		S_wI_w	0.1625	0.1625	0.1634	0.1685	0.1688
Wireless Indoor Localization	T-FRI	$S_{\bar{w}}I_{\bar{w}}$	0.1594	0.1704	0.1694	0.1775	0.1735
		S_wI_w	0.1592	0.1723	0.1709	0.1788	0.1742
	KH	$S_{\bar{w}}I_{\bar{w}}$	0.1695	0.1692	0.1706	0.1755	0.1741
		S_wI_w	0.1660	0.1695	0.1682	0.1762	0.1723
	CCL	$S_{\bar{w}}I_{\bar{w}}$	0.1639	0.1629	0.1610	0.1655	0.1668
		S_wI_w	0.1643	0.1640	0.1639	0.1699	0.1691
User Knowl- edge Modeling	T-FRI	$S_{\bar{w}}I_{\bar{w}}$	0.0265	0.0351	0.0300	0.0323	0.0309
		S_wI_w	0.0264	0.0346	0.0301	0.0338	0.0313
	KH	$S_{\bar{w}}I_{\bar{w}}$	0.0268	0.0305	0.0294	0.0322	0.0301
		S_wI_w	0.0268	0.0304	0.0292	0.0319	0.0302
	CCL	$S_{\bar{w}}I_{\bar{w}}$	0.0256	0.0312	0.0284	0.0306	0.0289
		S_wI_w	0.0256	0.0317	0.0281	0.0309	0.0292

et al., 2008] algorithms. The work introduces weights into rule antecedent attributes within these FRI procedures. The extensions have been systematically evaluated on ten benchmark classification problems, demonstrating the superior performance of these extended methods over their originals. Very importantly, as illustrated by the experimental analysis, the weighted FRI methods only require the least number (i.e., 2) of the nearest neighbouring rules to perform interpolation, thereby ensuring their efficiency in practical applications.

Such improved performances of the extended methods are attainable owing to the use of the relative significance degrees, or weights, of the individual rule antecedents to guide the selection of the nearest neighbouring rules for interpolation. These weights are derived from ranking attributes using the given sparse rule base only. The interpolation processes are modified by the weights as well, thereby reflecting different contributions made by different attributes in deriving the interpolated consequents. This differs from the existing approaches where all attributes are treated equally.

The ideas of weighted KH and weighted CCL essentially form two examples of the generalisation of the weighted T-FRI which was proposed in Chapter 4. Along with the weighted T-FRI, these three methods provide a choice for the alternatives to implement the weighted fuzzy interpolative reasoning framework. This offers more options to achieve more accurate performance for inference with sparse rule bases.

Chapter 7

Weighted Fuzzy Interpolative Reasoning for Interpretable Mammographic Mass Classification

THE proposed attribute weighted fuzzy rule interpolation (FRI) scheme has been seen successful applications in the field of classic pattern recognition, for tackling classification and prediction problems, as reported in Chapter 5. This chapter presents a systematic application of this scheme in the medical domain, for addressing the problem of mammographic mass classification (MMC).

The remainder of this chapter is organised as follows. Section 7.1 firstly introduces the background knowledge of MMC, motivating this application work. Section 7.2 describes the mammographic image data to be addressed. Section 7.3 describes the fuzzy rule-based interpolative reasoning system for MMC. Section 7.4 provides experimental evaluation of the implemented system, supported with statistical analysis. Finally, Section 7.5 presents a summary of this realistic application.

7.1 Preliminaries

This section presents a brief introduction to the problem of MMC, including a review of the relevant techniques for MMC and a discussion of challenging issues remaining to be resolved. The motivations for the current work are then reported.

7.1.1 Background

Breast cancer is one of the severest threats for women around the world. Early detection of breast lesions has been shown to provide an essential means to reduce the possibility of deterioration of patients' health conditions or even death. Amongst various tools available, mammography screening offers a particularly popular technique for identifying the presence of abnormalities in breasts. As a result, mammographic images are produced, in the form of films or more advanced recently, in that of full field digital mammograms, which are helpful to effectively detect and diagnose breast cancer by medical professionals.

Mass and microcalcification are two important early signs of abnormalities for detecting developing breast cancer, which are normally present in mammographic images. Masses are often indistinguishable from the surrounding parenchymal, resulting in more significant challenges for mass detection and classification. In general, an abnormal mass can be categorised into either benign or malignant. For instance, the standardised Breast Imaging Reporting and Data System (BI-RADS) [Samuels, 1998] characterises masses for determination of benign or malignant in terms of their shapes, margins and densities. This reflects how radiologists visualise the mammographic images for diagnosis. Benign masses are frequently found to be in round or oval shapes, having well-defined margins and low densities, whilst malignant masses are more likely in irregular shapes and have spicule margins with relatively high densities.

Reading mammograms is a very demanding task for radiologists, and the determination of whether an image shows a benign mass or malignant may be affected by the experience and subjective criteria of a certain radiologist who handles a given case. The development of Computer-Aided Diagnosis (CADx) techniques plays an effective supporting role in assisting medical professionals in the interpretation of medical images. Especially, a combination of using a CADx system and exploiting human expertise directly would greatly improve diagnostic accuracy and efficiency. A number of CADx systems have been studied and applied to support mammographic abnormality diagnosis (e.g., [Liu and Tang, 2013, Magna et al., 2016, Miranda and Felipe, 2015, Oliver et al., 2012, Pérez et al., 2015, Xie et al., 2016]). Most developed techniques can be referred to in the recent survey of such research in [Cheng et al., 2006, Oliver et al., 2010, Yassin et al., 2018].

Existing computational techniques may provide a second opinion for mammographic mass diagnosis, by dealing with the mammograms using pathological related knowledge. In general, most CADx systems for mammogram mass classification build their structures by following a number of key phases, including: image preprocessing, region of interest (ROI) extraction, mass segmentation, feature extraction and selection, and class determination. Various image features have been found in the literature for characterising mass properties, such as traditional features in terms of intensity, morphology, texture, etc. and features generated from advanced computational mechanisms like deep neutral networks [Wu et al., 2018]. Morphological (aka. geometric) features are one of the most common types used to discriminate mammographic masses [Pedro et al., 2019], typically extracted to represent the shape and boundary characteristics of masses. They are commonly adopted to support precise mass segmentation carried out by radiologists or CADx systems. This is because such features depict what radiologists visualise a mass lesion, which are essential to enable subsequent interpretation of the classification or diagnostic outcome.

From medical viewpoint, interpreting mammogram masses visually is a very demanding task for radiologists. It would therefore be a great assistance to be able to produce interpretable diagnoses from any CADx system in use. Recently, efforts have been made for improving accuracy of CADx systems for mammogram classification, such as those achieved by deep convolutional neural networks (DCNNs, e.g., [Lévy and Jain, 2016, Wu et al., 2018, Yi et al., 2017]), which have been seen to make great progress in meeting the visual recognition challenges. In such work, informative features are extracted to generate potential explanations for mammogram classification, by visually showing the edge of mass in saliency maps for example. However, to ensure interpretable feature representations requires the annotations of radiologists (or other alternative means) to correlate the DCNNs features with radiological features that reflect clinically relevant phenomena. This makes the interpretation progress and hence, the entire diagnostic system more complicated. It remains a difficult problem to discover clinically explainable interpretations for machine learning-based CADx systems.

7.1.2 Motivations

The question now is what intelligent classification methods can be better developed to facilitate the use of semantics-rich geometric mass features, in an effort to enhance

CADx systems' explainability explicitly. Fuzzy rule-based systems are known to be able to simulate human reasoning in decision support. Inference made by firing fuzzy if-then rules can be readily interpreted by human users. Such systems provide an effective tool to deal with the impreciseness and vagueness commonly incurred in real-world problems, including the description of mammographic mass characteristics. Fuzzy rule-based techniques therefore, have a natural appeal in establishing a CADx system for mammographic mass diagnosis. For example, Adaptive Neuro-Fuzzy Inference System (ANFIS) has been applied to classifying normal/abnormal mammograms, as well as to determining abnormal severity [Mousa et al., 2005]. Also, the classical Compositional Rule of Inference (CRI) [Zadeh, 1973] has been employed to perform mammogram diagnostic reasoning (e.g., classifying mammogram mass lesions into the well-known BI-RADS shape categories) [Miranda and Felipe, 2015, Vadivel and Surendiran, 2013].

Little work exists to explicitly interpret radiological phenomena of mass lesions in mammograms with the use of fuzzy rules, however. In addition, there may not be sufficient mammographic image data to enable the full exploitation of traditional fuzzy systems to perform required diagnostic tasks. As such, a fuzzy rule base inducted from the data may not cover the entire problem domain, resulting in the situations where certain observations can not match any of the rules in the rule base, thereby deriving no or wrong conclusions [Vadivel and Surendiran, 2013]. Fuzzy interpolative reasoning through fuzzy rule interpolation (FRI) can help to deal with exactly such sparse knowledge-based problems [Baranyi et al., 2004, Chen et al., 2015, Huang and Shen, 2006, Kóczy and Hirota, 1993a, Yang et al., 2017]. The efficacy of classical FRI techniques have been significantly strengthened with the recent advances in the literature, including the weighted FRI approach introduced earlier in this thesis, which no longer imposes the constraint that the rule antecedent features are of equal significance in decision-making. Instead, features are ranked with their relative weights exploited in the procedures of a conventional FRI method (e.g., the scale and move transformation-based FRI, T-FRI [Huang and Shen, 2008]). As demonstrated before, the resultant techniques have been successfully applied in tackling classification and prediction problems, inspiring the development reported herein.

Two key contributions to the relevant literature are to be reflected and reinforced in this application work: 1) an implemented fuzzy rule-based inference system for mass classification in mammograms, where fuzzy interpolative reasoning is embedded for the first time in a CADx system (for coping with sparse rule bases), supported by feature weight-guided FRI; and 2) an explicit explanation output from the CADx system, in the form of clinically interpretable rules using features of doamin semantics, thereby providing a "second opinion" for assisting radiologists to read mammograms.

7.2 Databases

The benchmark mammographic image datasets used in this work are adopted from the Breast Cancer Digital Repository (BCDR) [Lopez et al., 2012]. It is a wide-ranging and comprehensively annotated public database for mammographic disease study, especially for the development of breast cancer CADx techniques and for training medical physicians involved in the diagnostic, treatment or research of breast cancer and associated technologies. This repository is continuously being enriched and currently, contains cases of 1734 patients with mammography and ultrasound images, clinical history, lesion segmentation and selected pre-computed image-based descriptors.

BCDR consists of two different types of sub-repository: 1) a digitalised film mammography (FM)-based repository, and 2) a full field digital mammography (DM)-based repository. Both FM and DM repositories are divided into several sub-datasets including different number of cases, which form a common ground for fair comparison between various CADx systems for mammographic disease analysis. As with other established mammographic databases, digitalised film mammogram images have rather lower resolution whilst full field digital mammogram images are much more common nowadays (because of their higher spatial resolution and permitting more image manipulation to enable better visualisation). Without biases, the present work takes samples from both FM and DM sub-datasets, containing the following types of mass:

- BCDR-D01: comprised of 79 biopsy-proven lesions of 64 women, rendering 141 segmentations. All of them present suspicious mass, of which 85 are benign and 56 are malignant. Each image is a grey level mammogram in 14 bits with a resolution of 3328×4084 pixels.

- BCDR-F01: comprised of 200 biopsy-proven lesions of 190 women, rendering 362 segmentations, with mass lesions occurring in 231 segmented images where the number of benign and malignant masses are 112 and 119, respectively. Each image is a grey level digitalised mammogram in 8 bits with a resolution of 720×1168 pixels.

Note that multiple views of a single abnormality are involved in each sub-dataset, which results in the number of the abnormality segmentation being possibly more than that of the detected lesions (or the number of patients). This study exploits all views of a certain mass for conducting the evaluation of the proposed work, in order to demonstrate its potential practicality for constructing an interpretable fuzzy rule-based CADx system to classify mammographic mass lesions. As for the ultimate task of directly assisting the diagnosis of breast cancer, a pre-processing procedure for separating the mammographic images in different views needs to be taken into consideration.

Note that each mammogram image considered has a precise segmentation of identified lesion. In particular, the contour of mass is manually annotated by medical specialists. Fig. 7.1 shows examples of benign and malignant mass lesions with respective mass segmentations, taken from each of the two datasets.

7.3 Fuzzy Rule-based Interpolative Classifier

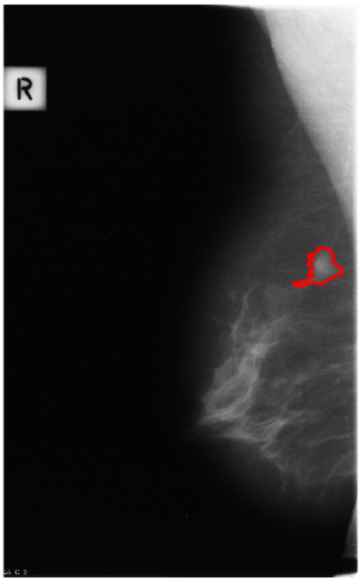
This section details the design and implementation of a rule-based system that works through the assistance of fuzzy interpolative reasoning, for classifying mammographic mass in mammogram images.

7.3.1 System Framework

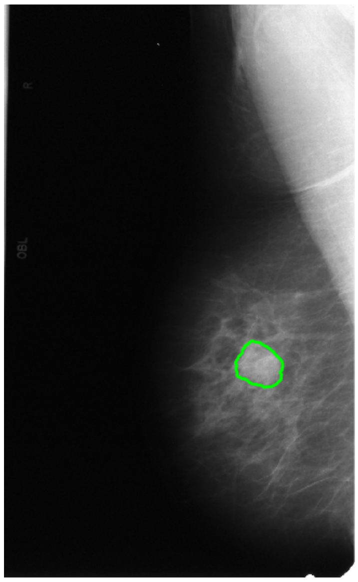
The workflow of the entire diagnostic system is specified as illustrated in Fig. 7.2. The general working process is as follows. Having identified a general region of interest (ROI) and segmented mass lesion from a given original mammogram image, a set of potentially descriptive features are extracted for characterising the properties

(a) Malignant mass in BCDR-D01

(b) Benign mass in BCDR-D01



(c) Malignant mass in BCDR-F01



(d) Benign mass in BCDR-F01

Figure 7.1: Samples of mass lesions with mass contours annotated.

of the image (particularly regarding the geometric shape, margin, density of mass lesion). The resulting mass features are evaluated by a feature ranking method, of two-fold objectives: 1) selection of more informative top features, and 2) assignment of weights to those selected ones in terms of their relative ranking scores. A fuzzy semantic rule base is generated from the given image database through the use of selected mass features as rule conditionals, by employing a certain standard fuzzy rule induction method (e.g., the one described in Appendix A).

Following the aforementioned preparation, the primary work for mass classi-

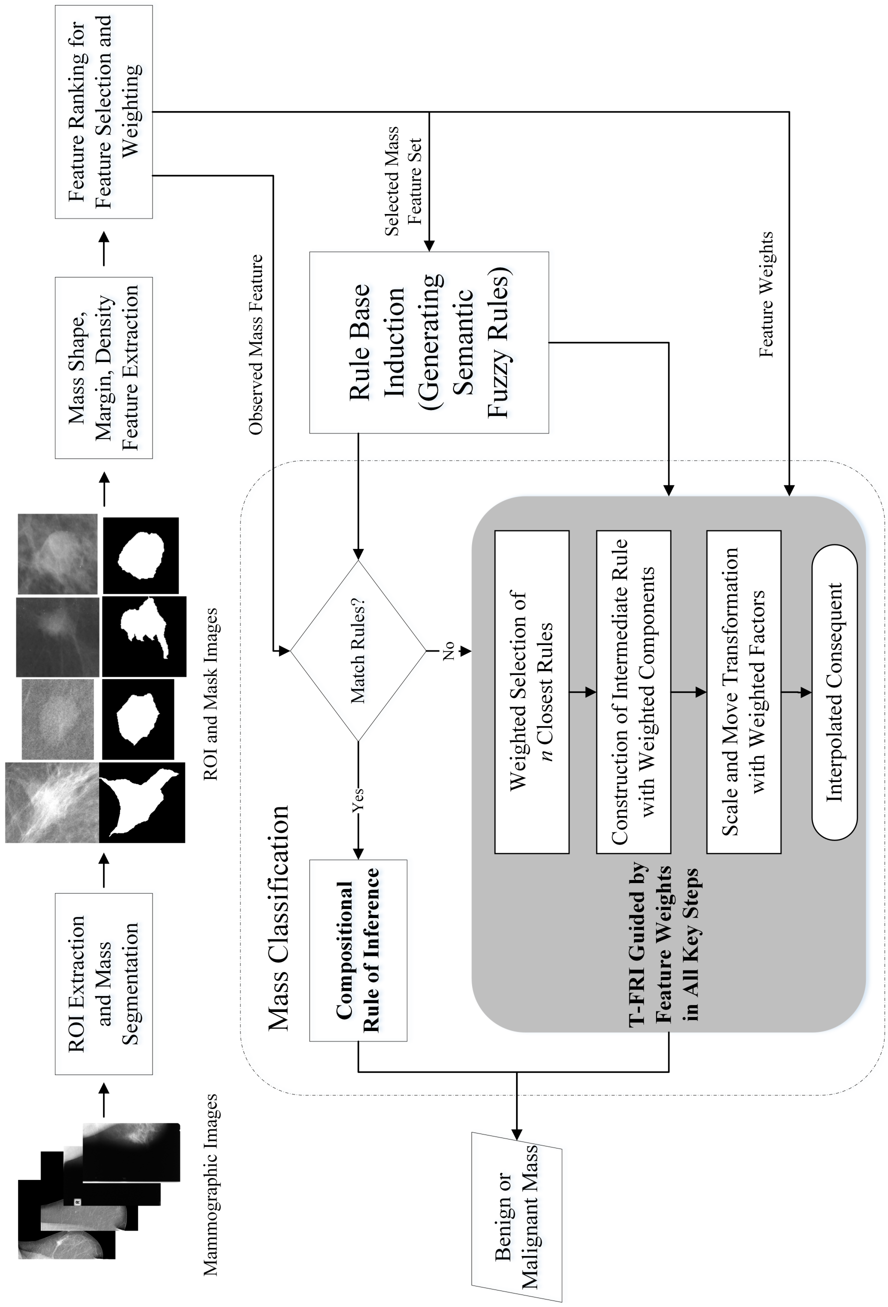
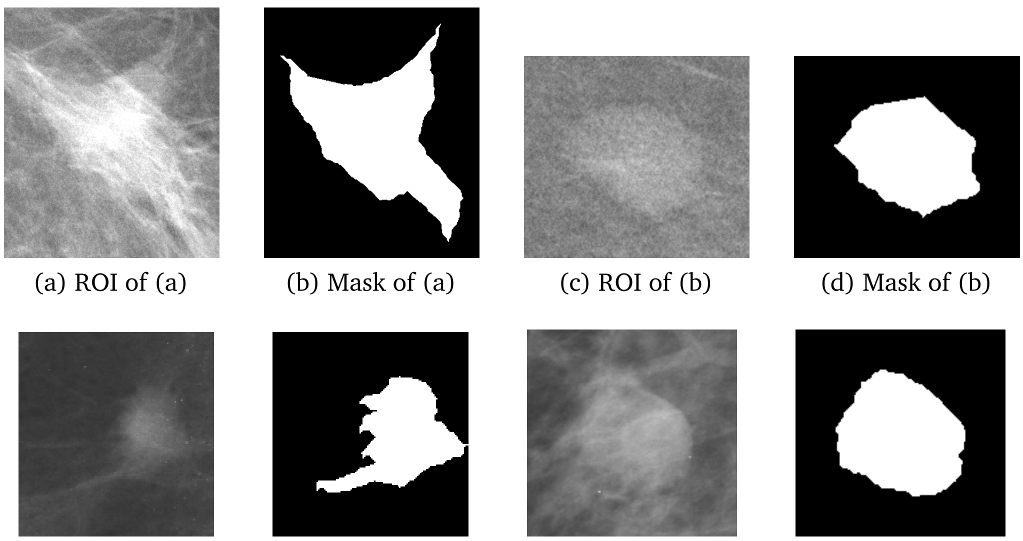


Figure 7.2: Fuzzy interpolative reasoning for mammographic mass classification.

fication is highlighted in the dashed box in Fig. 7.2. In particular, when a novel observed mass is present (represented with selected features) it is regarded as a new observation to be checked against the rules within the rule base. If it is matched by any existing rule, the rule is fired by the use of CRI. If there is no rule matching the observation, weighted fuzzy rule interpolation (where T-FRI is used for implementation here, though others such as those presented in Chapter 6 may be used as an alternative) to perform interpolative reasoning, estimating the benignancy or malignancy of the given mass. Technical details are provided in the following.

7.3.2 ROI Extraction and Mass Segmentation

In BCDR, each mammogram is associated with a precise segmentation of the underlying mass lesion. Since the focus of this work is on mass classification, the available contours of masses are adopted here for generating the ROI image and subsequently, the mass-segmented mask image of each given mammogram. These two images are chopped from the original mammogram, such that the observed mass locates in the centre. The resultant images consolidate the basis upon which to extract features in terms of mass shapes, margins and densities. Fig 7.3 shows examples of the ROI and mass-segmented mask images as per those mammogram samples displayed in Fig 7.1.



(a) ROI of (a) (b) Mask of (a) (c) ROI of (b) (d) Mask of (b)

(e) ROI of (c) (f) Mask of (c) (g) ROI of (d) (h) Mask of (d)

Figure 7.3: ROI and mass-segmented mask images of mass samples given in Fig. 7.1.

7.3.3 Mass Feature Extraction and Ranking

Given the ROI image and mass-segmented image of a mammogram, a set of features are extracted for characterising mass lesion in terms of the image properties such as mass shape, margin and density. Generally, the benign masses are frequently found to be in round or oval shapes, having well-defined margins and low densities, while the malignant masses are more likely in irregular shapes and have spicule margins with relatively high densities. Inspired by this observation, in this work, a total of 18 features are taken as the possible ones to distinguish benign and malignant masses, as listed in Table 7.1. This intuitive approach is based on the understanding of medical professionals practice, in that these two types of mass are often differentiated from their geometrical shape and boundary as well as density.

Benign and malignant masses may be found in rather different shapes. To reflect this viewpoint, six geometry features are extracted from the mask images of mass, including: mass area (F1), mass perimeter (F2), circularity measure (F3), convexity measure (F4), mass eccentricity (F5) and dispersion (F6). In particular, area and perimeter are basic shape descriptors for measuring the size of a mass. The features F3-F6 are metrics which define the morphological characteristics of masses in different shapes, potentially helpful to differentiate masses of regular shape from those of irregular, and to quantify the circularity and ellipticity of regular masses.

The margin of a mass offers another view for depicting the geometric properties of masses. Margin features can be grouped in two sub-categories. One is used to determine the degree of boundary roughness. Herein, five normalised radial length (NRL)-based statistical features (F7-F11) and compactness measure (F12) are employed to cover this aspect. The other group is to quantify the sharpness of margin intensity, with three margin gradient features (F13-F15) adopted to measure the pixel intensity variations over the boundaries of masses.

Mass shape and margin features characterise the morphological properties of mass regions, while the density features of mass reveal the intensity of mass region compared against its surrounding tissue. The last three features are therefore adopted to exploit the pixel intensity within a mass involved in the ROI images. In particular, the features F16 and F17 are computed with respect to the statistics relevant to the moments which measure the intensity of suspicious mass region. The contrast measure (F18) is the difference between the average grey level of the ROI and that of the surrounding region, evaluating the intensity variation within masses in contrast to that outside.

Table 7.1: Mass Features in Different Category for Characterising Mass Lesion

Mass Features		Physical Meaning
Shape	Area (F1) [Cheng et al., 2006, Dominguez and Nandi, 2008] Perimeter (F2) [Cheng et al., 2006, Xie et al., 2016]	Size Small values indicate small mass lesion
	Circularity (F3) [Cheng et al., 2006, Petrick et al., 1999]	Degree of roundness/circularity F3=1 for a circular mass and less than 1 for mass that departs from circularity
	Convexity (F4) [Cheng et al., 2006]	Relative amount that an object differs from a convex object F4=1 for convex mass (as with many benign masses) and less than 1 for nonconvex mass (as with many spiculated or malignant masses)
	Eccentricity (F5) [Dominguez and Nandi, 2008, Vadivel and Surendiran, 2013]	Degree of ellipticity Small values for circle-like ellipse and large for line segment-like ellipse
	Dispersion (F6) [Dominguez and Nandi, 2008, Vadivel and Surendiran, 2013]	Degree of irregularity (Density of region) Small values indicate regular masses while large values for irregular masses

Margin	Statistical normalised radial length (NRL) features (F7-F11) [Cheng et al., 2006, Petrick et al., 1999]: mean, SD, entropy, area ratio, zero-crossing count	Degree of boundary roughness
	Compactness (F12) [Dominguez and Nandi, 2008, Mu et al., 2008]	Small values indicate smooth contour (as with benign masses)
	Margin statistical gradient features (F13-F15) [Xie et al., 2016]: mean, SD, entropy	Intensity variations across the boundaries of mass Small values indicate flat edges while large values for sharp boundary
Density	Mass Intensity Mean (F16) [Cheng et al., 2006, Xie et al., 2016]	Average intensity value inside mass Small values indicate low density mass
	Mass Intensity Standard Deviation (F17) [Cheng et al., 2006]	Intensity variation inside mass Small values indicate little intensity variation within mass
	Contrast measure of ROIs (F18) [Cheng et al., 2006, Petrick et al., 1999]	Intensity variation between inside and outside of mass Small values indicate low density contrast

Note that there may exist redundant features among the extracted combinatorial properties of mass shape, margin and density. Obviously, such redundancy should be removed, not only to improve the performance of classifier (via the use of less features gaining efficiency and the reduction of measurement noise gaining effectiveness), but also to enhance interpretability of the diagnostic system (with less complex rules). In this chapter, a feature ranking mechanism taken from the core of the popular Relief-F algorithm [Kononenko, 1994] is employed to evaluate individual mass features. This differs from the work in Chapter 5, where information gain was adopted. This is purely for the purpose of demonstrating the diversity of feature evaluation methods for possible use. As stated before, other evaluation algorithms may also be utilised as an alternative if preferred.

The use of the feature ranking mechanism results in a set of scores that indicate the relative importance of each feature in the determination of benign and malignant mass. Intuitively, those features which have relatively lower scores may have poorer capability in the discrimination of different classes, and thus a subset of features are selected whose score values are higher than the average. The average score is herein utilised in order to ensure the process is automated; otherwise, if desirable, a pre-defined threshold may be used for the removal of low-ranking features.

Without losing generality, suppose that there are m features being selected, each of which has a ranking score $RS_i, i = 1, 2, \ldots, m$. These different score values can then be normalised as weights associated with each of the selected individual features, as follows:

$$W_i = \frac{RS_i}{\sum_{t=1,\ldots,m} RS_t} \tag{7.1}$$

Given their underlying definition, the resulting normalised ranking scores have a natural appeal to be interpreted as the relative significance degrees of the contribution that a remaining feature may make to the decision, regarding the benignancy or malignancy of the mass. Such weights will also be utilised to guide the fuzzy rule-based interpolative inference system for mass classification as to be discussed later.

7.3.4 Generation of Fuzzy Classification Rules

Having represented mass lesions in mammograms with selected mass shape, margin and density features, fuzzy rules for mass classification can be generated from given images whose decision classes are known. More specifically, the fuzzy rule base for mass classification consists of fuzzy *if-then* rules whose antecedent attributes are selected mass features and consequent attribute is the corresponding mass lesion type (i.e., Benign or Malignant).

The fuzzy values for each antecedent feature are fuzzified linguistic terms, which are defined in terms of the physical meaning of the underlying mass features (that are given in Table 7.1). Different values of the numerical metrics defining the features indicate different properties of a certain mass (including: shape, margin and density). Generally, the linguistic terms describing the features can be given in order, such as "*..., Small ,..., Medium ,..., Large ,...*". Table 7.2 lists the linguistic values used in this work, mimicking the terms used by the medical professionals in the field concerned. The number of linguistic terms adopted by each mass feature can be determined from data by a data-driven method (see later for evaluation). From this definition, a fuzzy rule base is inducted from the extracted feature data, by promoting any hype-grid delimited by the fuzzy feature values that is hit by at least one given data. Note that any standard fuzzy rule induction method may be employed to create the rules, which is not the focus of this work. Unless stated otherwise, rules are herein learned from the selected mass features based on the use of the classical method of [Wang and Mendel, 1992], a summary of which is given in Appendix A.

A possible rule, for example, from the learned rule base may be represented such that

If *Area* is *Small* and ... *Circularity* is *Large* and ... *NRL zero-crossing* is *Small* and *Margin gradient mean* is *Large* and ... *Density contrast* is *Small*, then *Mass* is *Benign*.

From the underlying semantics of the morphological and density features, this rule can be directly mapped onto the following, using the linguistic terms given in Table 7.2:

If *Mass* is *Small* and ... *Mass shape* is *Very Like Circular* and ... *Mass margin* is *Smooth* and *Margin* is *Circumscribed* and ... *Mass density contrast (between inside and outside mass)* is *Low*, then *Mass* is *Benign*.

Using fuzzy rules like the above helps facilitate the understanding of any conclusion drawn regarding whether a new mammogram stands for a benign or malignant mass, through the use of the fuzzy interpolative reasoning system as described next.

Table 7.2: Gradual Linguistic Terms Defined for Mass Features as Numerical Values Vary from *Small* to *Large*

Mass Features		Linguistic Terms
Shape	Area (F1) Perimeter (F2)	Small,..., Medium,..., Large (mass)
	Circularity (F3)	Unlike,..., Less Like,..., Very Like (circular mass)
	Convexity (F4)	Unlike,..., Less Like,..., Very Like (convex regular mass)
	Eccentricity (F5)	Very Like,..., Less Like,..., Unlike (oval mass)
	Dispersion (F6)	Regular,..., Less Regular,..., Irregular (mass)
Margin	Statistical normalised radial length (NRL) features (F7-F11) Compactness (F12)	Smooth,..., Slight undulated,..., Irregular (contours)
	Margin statistical gradient features (F13-F15)	Obscured,..., Blurred,..., Circumscribed (Well-/ Sharply-Defined) (margins)
Density	Mass Intensity Mean (F16)	Low,..., Isodense,..., High (mass density)
	Mass Intensity Standard Deviation (F17)	Low,..., Medium,..., High (mass density variation)
	Contrast measure of ROIs (F18)	Low,..., Isodense,..., High (density contrast)

7.3.5 Feature Weight-Guided Interpolative Reasoning

When a new observation is present, in terms of a set of measured feature values (representing an unknown mass lesion), all rules in the rule base are used to match against it in order to derive a diagnostic conclusion. However, the rule base learned from previously given data may be sparse, especially when only limited source data (or classified medical mammographic images) are available. Thus, checking against all the available rules cannot fully cover the entire problem domain. That is, there exist situations where no rules can be found that match the new observation, leading to no conclusion to be drawn. To enable approximate inference on the unmatched observation, FRI is utilised. Thus, the previously developed feature weighted FRI is adapted to implement the required interpolative reasoning for mass classification.

The formal illustration of the feature weighted FRI approach can be referred to technical details presented in Chapter 4. In particular, suppose that each antecedent attribute in the learned (sparse) fuzzy rule base is now associated with a weight through feature ranking as given in Section 7.3.3. For simplicity and consistency throughout this thesis, triangular membership functions are employed for implementing the CADx system, with each fuzzy set represented by three characteristic points, as previously illustrated in Section 2.1 of Chapter 2.

In this work, the number of closest rules, n is set to 2 for conducting rule interpolation. This follows the empirical conclusion drawn previously, in that the adoption of the least number of closest rules (i.e., $n = 2$) is able to achieve a superior performance for feature weighted T-FRI. Such a set up normally has a high classification accuracy while saving computational costs.

Once the weighted FRI is reached, when a sparse rule base is learned from source data and a novel observation finds no rules to match, the required consequent can be derived. The entire interpolative process is guided by the feature weights. Note that for those matched observations, the classification results are directly obtained by firing the matched rules without going through interpolation. As with many fuzzy rule-based systems, the resultant consequent fuzzy sets are required to be defuzzified for providing a class label, returning the conclusion on classification. Obviously, in the present CADx system, the conclusion drawn over the given mass is whether its type is benignancy or malignancy.

Finally, to reinforce the understanding and to help implement the proposed mammographic diagnostic system, Algorithms 8 and 9 present the pseudocode for the training and application (or testing) of the classification system, respectively. They jointly reflect the overall system framework as illustrated in Fig. 7.2. Note that the subroutine implementing the core shaded part of Fig. 7.2, i.e., the procedure for feature weight-guided interpolation, is simply outlined in Line 8 of Alg. 9 without comprehensively detailing it. This is because the work presented herein is aimed to offer a practical application of weighted FRI, detailed pseudocode of which can be found in Alg. 7 in Section 5.1 (i.e., Weighted T-FRI $B^* = WeightedTFRI(R, o^*, n, W)$ where $n = 2$).

Algorithm 8 Pseudodode of Mammographic Mass Classifier under Training

Input:

- Training dataset with mammographic images labelled with mass type

Output:

- Selected conditional attributes and their relative weights
- Rule base

Identify mass ROI images for each of input mammographic images;
Segment mass aided with available contours provided in dataset, resulting in mass-segmented mask images;
Extract K mass shape (F1-F6), margin (F7-F15) and density (F16-F18) features ($K = 18$, as specified in Table 7.1) for each mammogram using pairs of ROI and mass-segmented mask images;
Rank extracted mass features (F1-F18) of training dataset to obtain ranking score $RS_i, i = 1, 2, \ldots, K$;
Select top m features $F = \{RS_i, i = 1, \ldots, m\}$ such that $RS_i > \frac{1}{K}\sum_{t=1}^{K} RS_t$;
Calculate feature weights W in terms of Eqn. (7.1);
Generate fuzzy rule base R using selected mass features and mass types;
Return F, W and R

7.4 Experimental Evaluation

This section presents a systematic experimental evaluation of the proposed fuzzy rule-based interpolative system for mammographic mass classification. The results are reported on the classification accuracy, sensitivity, specificity, the area under the Receiver Operating Characteristic (ROC) curve, and the ratios of false positives and false negatives over the size of the testing data. These are supported by running

Algorithm 9 Pseudocode of Mammographic Mass Classifier in Action

Input:

- Rule base (R) generated from training
- Selected features (F) and their relative weights (W) produced from training
- Unknown mammogram to be classified

Output:

- Mass category (i.e., benignancy or malignancy)

Identify mass ROI image of given mammogram;
Segment mass, resulting in mass-segmented mask images;
Extract $|F|$ features (as specified in F, where $|F|$ stands for F's cardinality), serving as observation o^* to be classified;
Match o^* against each rule in rule base R;
if *matched with at least one rule* **then**
  Fire matched rule(s) using CRI to obtain required consequent B^* for o^*;
else
  Execute weighted FRI to compute $B^* = WeightedTFRI(R, o^*, 2, W)$;
end if
Defuzzify B^* as a class label;
Return Benign or Malignant mass

nonparametric Wilcoxon signed-rank tests for validating the statistical significance of the classification performance.

7.4.1 Experimental Setup

To have a common ground for fair comparison, all of the given mammographic images which contain mass lesions provided in BCDR-D01 and BCDR-F01 datasets are employed to conduct the evaluation, respectively. The experimentation is carried out independently over these two datsets for full field digital mammograms and digitalised film mammograms. As indicated previously, the mass contours annotated by medical specialists are used for the generation of mass-segmented mask images, where the distance between the margin of the chopped box and the provided mass boundary is empirically set as 30 pixels. The corresponding ROI images are of the same size as that of the mask images, while each sharing the same location as their respective original. Again, examples of those can be found in Fig. 7.3 of Section 7.3.2.

The classification performance is herein evaluated by 10-fold cross validation randomly repeated for 10 times for both datasets. The partition of each antecedent attribute domain (which is normalised) into triangular membership fuzzy values is achieved by approximating what is learned by the use of Fuzzy C-Means (FCM) [Bezdek et al., 1984]. The number of triangular membership functions (i.e., clusters) for each attribute tuned by FCM is determined by the standard method of [Chen and Wang, 1999].

Comparative experimental studies are carried out for classifying mammographic masses, amongst the following three situations: 1) matching the rules in the learned rule base using CRI only for classification (as per classical fuzzy inference systems without FRI), 2) performing CRI for those matched testing observations and conventional unweighted T-FRI for those unmatched ones, and 3) running CRI for matched rule firing and weighted T-FRI for interpolative rule-based classification.

To comprehensively evaluate the classification performance, the following four commonly used metrics are adopted: classification accuracy, sensitivity, specificity and AUC (i.e., area under ROC curve). These performance indices are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7.2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7.3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7.4}$$

where TP, FP, TN and FN stand for the number of: true positives, false positives, true negatives and false negatives, respectively. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings and then, the area encompassed by the plotted curve is computed. All of the four evaluation metrics take values between 0 and 1, and a good diagnostic test is obtained when these are close to 1. In addition, two ratio-based performance criteria are also checked, namely FP ratio and FN ratio, which are defined as the ratio between the number of FP over the data size, and that between the number of FN over the data size, respectively. Here, data size stands for the number of images tested. These two ratios are computed as follows:

$$\text{FP ratio} = \frac{FP}{\text{Number of testing images}} \tag{7.5}$$

$$\text{FN ratio} = \frac{FN}{\text{Number of testing images}} \tag{7.6}$$

Smaller values of these ratios indicate better classification, of course.

7.4.2 Results and Discussion

Comparative experimental results are reported and discussed in this section, including aspects regarding classification interpretability as well as performance measurements.

7.4.2.1 Interpretability of Fuzzy Rules for Diagnosis

The mechanism for mammographic mass classification is achieved by the use of semantic fuzzy rules, through rule firing for novel observations that match a certain given rule or rule interpolation for those that match no rules. As indicated before, such fuzzy rules are human interpretable because of the employment of selected semantics-rich, morphological and density features as rule antecedent attributes. In the following, two examples are provided to show the interpretable diagnostic procedure of rule matching (i.e., CRI) and that of rule interpolation for mass classification, respectively.

A. Running CRI over Matching Rule(s)

For BCDR-D01 dataset, nine top-ranked features are selected to generate the fuzzy rule base. These are: Perimeter (F2), NRL entropy (F9), Mass intensity standard deviation (F17), Margin gradient entropy (F15), Compactness (F12), Mass intensity mean (F16), Margin gradient SD (F14), Convexity (F4), Margin gradient mean (F13). All types of mass feature are involved. In particular, F2, F4 and F17, F16 are mass shape and density features, respectively, while the remaining are mass margin descriptors. These features utilise 3, 3, 4, 4, 3, 3, 3, 4, and 4 fuzzy membership sets representing the underlying linguistic terms, returned by the application of FCM. In particular, the terms used for three-membership features are "*Small, Medium, Large*", and those for four-membership features are "*Small, Medium-small, Medium, Large*" or "*Small, Medium, Medium-large, Large*", depending on which end over the normalised interval [0,1] the partitions are closer to.

Consider as an illustrative example, the observation consisting of the following feature values:

[F2, F9, F17, F15, F12, F16, F14, F4, F13] =

[0.0760, 0.3178, 0.0368, 0.4178, 0.2690, 0.1181, 0.0709, 0.9179, 0.0865]

with the original mammogram, mass-segmented mask and ROI images shown in Fig. 7.4. There are four fuzzy rules in total which match this observation, of which the one shown below has the largest matching degree:

If *Perimeter (F2)* is *Small* and *NRL entropy (F9)* is *Medium* and *Mass intensity SD (F17)* is *Small* and *Margin gradient entropy (F15)* is *Medium* and *Compactness (F12)* is *Small* and *Mass intensity mean (F16)* is *Small* and *Margin gradient SD (F14)* is *Small* and *Convexity (F4)* is *Large* and *Margin gradient mean (F13)* is *Small*, then *Mass* is *Benign*.

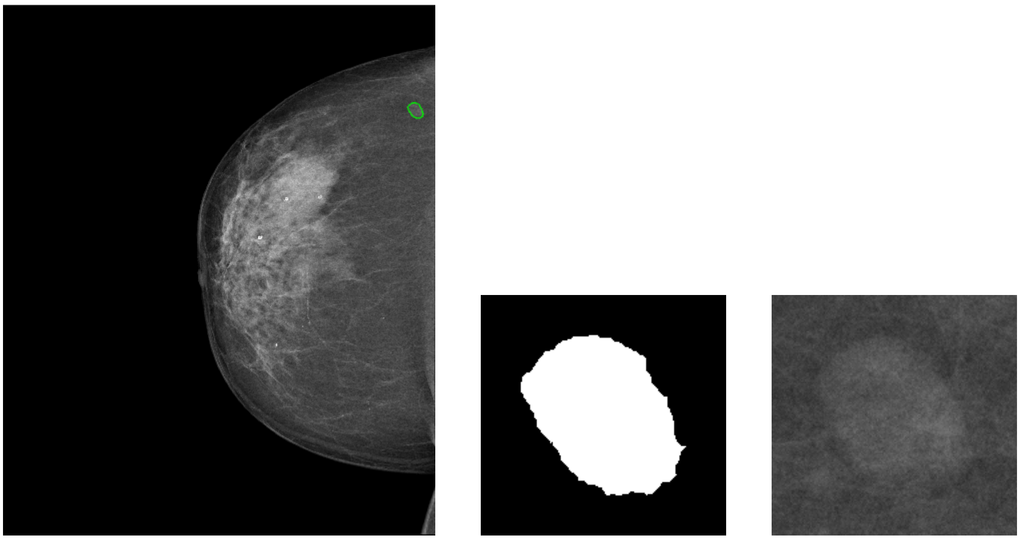
Considering the semantic meaning of each feature given in Table 7.2, the above rule can be directly translated into:

If *Mass size* is *Small* and *Mass contour* is *Smooth and Blurred* and *Mass density (and its variation)* is *Low* and *Mass* is *Very Like a convex regular region*, then *Mass* is *Benign*.

Firing this rule successfully classifies the mass as Benign, as shown in Fig. 7.4. It visually recognises the mass lesion in terms of its geometrical shape, contour and density properties, which can be readily understood by medical specialists or explained to the patient.

B. Running Weighted Rule Interpolation due to No Matching Rules

As illustrated in Fig. 7.2 of Section 7.3.1, feature weight-guided FRI procedure is triggered by any observation that matches no rules in the sparse rule base, deriving an interpolative classification of the mass category. In this case, selecting two closest neighbouring rules forms the starting point and sets the foundation for rule interpolation.



(a) Original mammogram (b) Mass-segmented mask (c) Chopped ROI

Figure 7.4: Benign mass classified by matched fuzzy rules.

As with the case for BCDR-D01, in BCDR-F01, all extracted features are ranked first, resulting in the top six being selected. These are: Compactness (F12), Convexity (F4), Circularity (F3), NRL entropy (F9), NRL zero-crossing count (F11) and Mass intensity mean (F16). In particular, F4 and F3 are mass shape features, F12, F9 and F11 are mass margin features, and F16 is selected again as the density descriptor in this dataset.

The number of fuzzy membership functions learned for these selected features are 4, 4, 5, 4, 2, 4, respectively. The fuzzy terms taken by the four-membership features attain the same rule as set in BCDR-D01, while the (only) two-membership feature has two alternatives (i.e., "*Small, Large*") and the remaining one has five fuzzy values, taking from "*Small, Medium-small, Medium, Medium-large, Large*".

Consider the case where the following observation is given which has no rules matched:

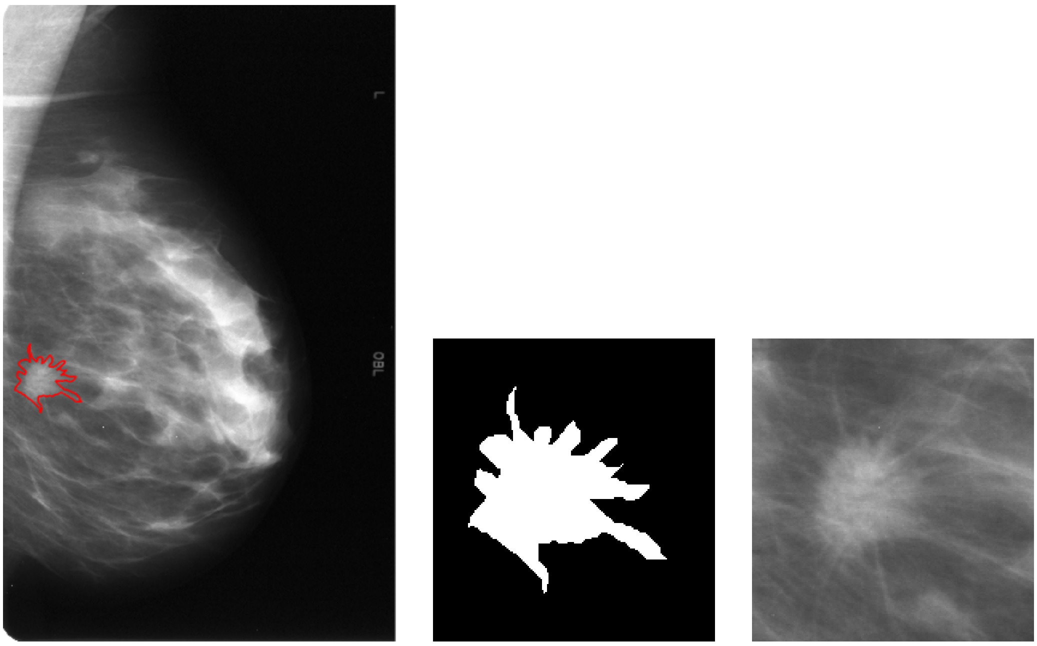
[F12, F4, F3, F9, F11, F16] = [0.9184, 0.2868, 0.2456, 0.8442, 0.4595, 0.4882]

The original mammogram, mass-segmented mask and ROI images for this case are shown in Fig. 7.5. From this, two fuzzy rules are found to be the closest to the given observation, which are:

Rule 1: If *Compactness (F12)* is *Large* and *Convexity (F4)* is *Medium-small* and *Circularity (F3)* is *Medium-small* and *NRL entropy (F9)* is *Medium-large* and *NRL zero-crossing count (F11)* is *Large* and *Mass intensity mean (F16)* is *Medium*, then *Mass* is *Malignant*.

Rule 2: If *Compactness (F12)* is *Medium-large* and *Convexity (F4)* is *Medium* and *Circularity (F3)* is *Medium* and *NRL entropy (F9)* is *Medium-large* and *NRL zero-crossing count (F11)* is *Large* and *Mass intensity mean (F16)* is *Medium*, then *Mass* is *Malignant*.

Both rules give malignancy as the conclusion. Having taken into account the semantic linguistic values used for each mass feature in Table 7.2, these two selected rules jointly lead to the following interpolated rule, with detailed computational process omitted to save space:



(a) Original mammogram (b) Mass-segmented mask (c) Chopped ROI

Figure 7.5: Malignant mass classified by feature weight-guided FRI.

If *Mass* is *Less Like a circular regular region* and *Mass contour* is *Irregular* and *Mass density* is *Slightly high*, then *Mass* is *Malignant*.

The final interpolated consequent also indicates malignancy for the observed mass. As can be seen, classifying mammographic mass through interpolating two semantic fuzzy rules offers clear interpretability.

Collectively, the interpretability of the proposed fuzzy rule-based diagnostic system is shown by the process of inferring the category of mammogram mass, running either CRI over matched rule(s) against a given observation or weighted rule interpolation when there is no matching rule. Such interpretability is empowered by the employment of selected semantics-rich, morphological and density features as rule antecedent attributes, in conjunction with the underlying logic relationships between these attributes and the classification outcome. Only clinically explainable fuzzy rules are used for classification. This forms a significant contrast with existing techniques for addressing the problem of mass classification. For instance, in attempting to building an interpretable CADx system using a deep convolution neural network (DCNN)-based framework, such as DeepMiner [Wu et al., 2018], great effort has been devolved to discovering interpretable representations in deep neural networks so as to provide explanations for medical predictions. Unfortunately, generation of explanations for DCNN-based mammogram classification requires sophisticated expert annotation regarding any interpretable network units. Another attempt is to reveal visually interpretable images extracted from a DCNN, being only concerned with the edge of masses in saliency maps [Lévy and Jain, 2016]. Yet, no human-like linguistic explanation is produced automatically, unlike what is facilitated in the present rule-based approach.

7.4.2.2 Performance Based on Fairly Dense Rule Base

In this part of investigation, all fuzzy rules in the learned rule base are used for mammographic mass classification. Table 7.3 shows the results with respect to the six performance criteria, namely classification accuracies, sensitivities, specificities, AUC, FP ratio and FN ratio, which are obtained by averaging the outcomes of 10×10-fold cross validation. In particular, results on the row named CRI are those achieved by firing matched rules only, those on T-FRI by aiding CRI with classical T-FRI, those on W-T-FRI by combining CRI and feature weighted T-FRI. The classification outcome is obviously unknown for cases where CRI is used alone to deal with any unmatched observation, in which case an error is recorded while calculating the accuracy, sensitivity, specificity, FP ratio and FN ratio, but this does not apply to the computation of AUC.

The performance of CRI provides the baseline for comparison. As can be seen in Table 7.3, most of the testing samples are matched with the learned rule(s), resulting

Table 7.3: Performance on BCDR-D01 and BCDR-F01 by 10×10 Cross Validation

BCDR-D01						
Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
CRI	83.44	78.57	86.59	-	0.0814	0.0857
T-FRI	91.22	87.85	93.41	0.9607	0.0400	0.0485
W-T-FRI	91.65	88.93	93.41	0.9614	0.0400	0.0442
BCDR-F01						
Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
CRI	83.73	81.30	86.27	-	0.0668	0.0967
T-FRI	84.28	82.14	86.49	0.9019	0.0657	0.0924
W-T-FRI	84.28	82.14	86.49	0.9023	0.0657	0.0924

in reasonable classified results for both datasets. This is not surprising as the datasets used for training have been fairly comprehensive. Nevertheless, the rule base is not complete, there are uncovered problem spaces for which T-FRI and W-T-FRI can help improve the performance. Indeed, the use of either FRI method significantly strengthens the effectiveness of CRI on BCDR-D01, in terms of the improvement on classification accuracy, sensitivity and specificity, and in the reduction of both false positive ratio and false negative ratio. This shows the potential of fuzzy interpolative inference for coping with challenging situations where the given rule base fails to include rules matching a novel observation.

Applying the feature weighted FRI method has shown a slight further enhancement over the use of the popular T-FRI. The statistical significance is herein verified by Wilcoxon signed-rank test (with the parameter $p = 0.0312$). This demonstrates that the best AUC performance is attained by the use of W-T-FRI for both datasets, with 0.9614 and 0.9023 for the two datasets, respectively. This performance is comparable to that of the state-of-the-art CADx systems for mammographic mass classification, where the recorded best AUC measures are 0.9650 and 0.8940, respectively for BCDR-D01 and BCDR-F01 (see [Moura and López, 2013, Moura et al., 2013]). Yet, the classification process, and hence, the results of running the existing methods are not so easy to interpret as their counterparts of the proposed approach implemented herein. More importantly, the performance improvement becomes much more significant when considering situations where only a sparse rule base is available, as to be shown next.

7.4.2.3 Performance Based on Very Sparse Rule Base

The classification results presented in the preceding part of experimental evaluation are achieved by the use of the entire rule base learned from the data available. This is the situation that a real-world application would encounter. Even for the examined problem where a good amount of training data is exploited to generate a fairly dense rule base, as with the investigated case, sparseness may exist. This itself already shows the need for the employment of FRI techniques. However, there are practical situations where not sufficient training data is obtainable, especially when dealing with certain novel medical cases. It is therefore very interesting to investigate how the T-FRI in general and the W-T-FRI method in particular may bring forward any benefits in such situations. For this purpose, without complicating the experimental studies by introducing different datasets, here, two rule bases which are much sparser than the one used previously are artificially generated, by randomly removing a number of learned rules from the originally used. Note that this artificially imposed removal is for academic investigation only; in real application, unless there is inconsistency or redundancy, learned rules are not to be removed.

Table 7.4 shows the averaged results of this investigation, in relation to the percentage of rules removed. Particularly, the two sparser rule bases run are created by randomly deleting 30% and 70% of the learned rules, respectively. As expected, and reflected by this table, the performance of applying CRI alone declines dramatically as the proportion of rules remaining available decreases. The accuracies drop 30.01% (=83.44%-53.43%) and 60% (=83.44%-23.44%) for BCDR-D01 and 45.31% (=83.73%-38.42%) and 67.4% (=83.73%-16.33%) for BCDR-F01, respectively. The resultant performance deteriorates so much that such an approach is no longer acceptable in practice.

On the contrast, both FRI methods have shown to be able to alleviate such performance decline. With the employment of a FRI mechanism present CADx system maintains a strong capability in distinguishing suspicious mass lesions when CRI performs poorly, given only a considerably sparse rule base. Even when just a small proportion of rules remains available (for the cases where 70% of the rules are removed), the classification performance (regarding accuracies, sensitivities and specificities) is still at an approximate rate of 80% on the BCDR-D01 dataset and at high 60% on BCDR-F01. Regarding the FP and FN ratios, a significant reduction in these for both datasets has been shown by the use of either T-FRI or W-T-FRI as compared to the use of just CRI. Together, these results strongly demonstrate the significant effectiveness of fuzzy interpolative reasoning for resolving the problems involving a sparse rule base.

Examining more closely by comparing the performance of T-FRI and that of W-T-FRI, as the rule base is reduced to be much sparser, the improvement of W-T-FRI over T-FRI becomes more notable. In particular, the classification accuracy is enhanced by 2.36% and 4.08% with regards to 30% and 70% reduction of the rules on BCDR-D01, and by 1.62% and 3.73% on BCDR-F01. Furthermore, Table 7.5 summarises the average number of testing samples that require rule interpolation in each of the three inference situations (namely, running CRI alone, and CRI with T-FRI or W-T-FRI).

Table 7.4: Performance Based on Sparse Rule Base by 10×10 Cross Validation

Dataset	Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
BCDR-D01		Sparser Rule Base 1 (30% removed)					
	Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
	CRI	53.43	44.04	59.60	-	0.2452	0.2238
	T-FRI	83.71	78.57	87.06	0.8918	0.0786	0.0857
	W-T-FRI	86.07	81.55	89.02	0.9010	0.0666	0.0738
		Sparser Rule Base 2 (70% removed)					
	Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
	CRI	23.44	12.95	30.29	-	0.4232	0.3482
	T-FRI	77.67	74.55	79.70	0.8589	0.1232	0.1018
	W-T-FRI	81.75	79.02	83.53	0.8784	0.1000	0.0839
BCDR-F01		Sparser Rule Base 1 (30% removed)					
	Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
	CRI	38.42	29.41	47.99	-	0.2532	0.3652
	T-FRI	73.47	71.22	75.89	0.7948	0.1174	0.1489
	W-T-FRI	75.09	73.95	76.34	0.8238	0.1152	0.1347
		Sparser Rule Base 2 (70% removed)					
	Schemes	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	FP ratio	FN ratio
	CRI	16.33	12.60	20.24	-	0.3884	0.4521
	T-FRI	62.60	63.30	61.91	0.6833	0.1855	0.1898
	W-T-FRI	66.33	68.06	64.58	0.7040	0.1724	0.1652

Note that RB in this table and the next, stands for Rule Base. It is evident that the more unmatched samples, the more opportunities there are for the FRI methods to perform.

Table 7.5: Average Number of Testing Samples

Dataset	Number of Samples Requiring Interpolation/Total (per Fold)		
	Fairly Dense RB	Sparser RB 1	Sparser RB 2
BCDR-D01	1.24/14	5.85/14	10.50/14
BCDR-F01	0.21/23	12.77/23	18.70/23

Comparative performance is also measured through ROC analysis. The ROC curves resulting from running the standard T-FRI and feature weighted T-FRI over the use of different rule bases are given in Fig. 7.6, on both BCDR-D01 and BCDR-F01. Whilst it is not surprising that the best performance is achieved using the fairly dense rule base for both methods, W-T-FRI is shown to be less sensitive to the deterioration of sparsity of the rule base.

Last but not least, as indicated previously, to further determine the statistical significance in performance improvement of T-FRI over CRI, and that of W-T-FRI over T-FRI, the nonparametric Wilcoxon signed-rank tests are conducted. This is carried out for the classification accuracies obtained from the use of three different inference mechanisms implemented, with three different sparsities of the rule base on both datasets. Table 7.6 lists the p-value of each pairwise test. As can be seen in this table, all but one expectable test show relative small p-values (e.g., $p < 0.05$), which reflects the statistical significance of outperformance in each comparison. The only exception (with $p = 1$) is for the case comparing W-T-FRI against T-FRI on BCDR-F01 when the originally learned, fairly dense rule base is employed.

Table 7.6: P-value in Statistical Wilcoxon Signed Rank Test

	BCDR-D01		
	Original RB	30% Reduced RB	70% Reduced RB
CRI vs. T-FRI	5.65×10^{-13}	1.34×10^{-11}	3.31×10^{-8}
T-FRI vs. W-T-FRI	0.0312	8.03×10^{-5}	1.71×10^{-4}
	BCDR-F01		
	Original RB	30% Reduced RB	70% Reduced RB
CRI vs. T-FRI	0.0019	3.36×10^{-8}	1.56×10^{-6}
T-FRI vs. W-T-FRI	1	0.0169	4.88×10^{-4}

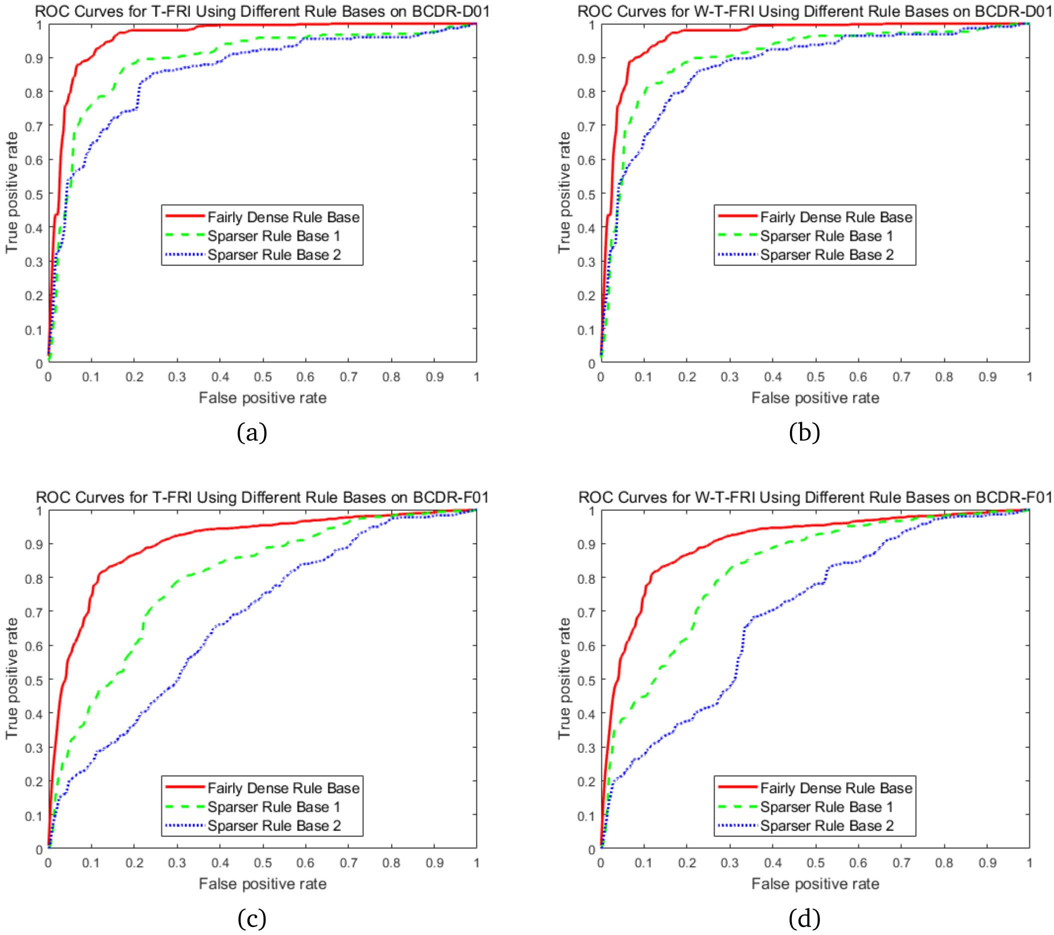


Figure 7.6: ROC for T-FRI and W-T-FRI using rule bases of different sparsity.

7.5 Summary

In this chapter, a novel fuzzy rule-based diagnostic system for mammographic mass classification (MMC) has been presented. The system is able to derive a conclusion for unknown observed masses that have no rules to match. The diagnostic outcomes are interpretable as the rules are learned over selected features in terms of mass geometric and density properties, with feature values represented in linguistic terms. The effectiveness of adapting feature weighted fuzzy rule interpolation as the core of the implemented system has been systematically evaluated and demonstrated, capable of dealing with rather sparse rule bases. This has been accomplished through comparison with the state-of-the-art work on mammogram datasets.

The applied feature weighted FRI presented in this chapter is essentially a realistic application of the attribute weight-guided FRI that was proposed in Chapter 4, through integrating weighted FRI into a realistic MMC diagnostic system. The implemented system (as shown in Fig. 7.2 and specified in Alg. 8 and Alg. 9) may seem a bit inconsistent when compared against the general workflow of weighted T-FRI (see Fig. 4.3 in Chapter 4), in terms of the derivation of the feature weights using feature ranking. The system here has used the weights from feature ranking via training data, while the generic framework is supported to apply the weights learned from the originally unweighted sparse rules. This is simply because the starting point of the general framework is a given sparse rule base, which is not the case for the present problem. Nonetheless, how the weights are obtained makes no difference for the weighted FRI method to work. This also provides a case to show the flexibility of the proposed weighted fuzzy interpolative reasoning.

Chapter 8

Conclusion

THIS chapter concludes the thesis. Firstly, a summary of the research presented in this thesis is given, which also re-states the contributions made from the study. Secondly, possible future work is outlined, including several further developments for the proposed innovative reverse engineering mechanism as well as those for the weighted fuzzy interpolative reasoning techniques.

8.1 Thesis Summary

The core work presented in this thesis is a novel fuzzy rule interpolation (FRI) approach that significantly reinforces the power of fuzzy interpolative reasoning. It works by exploiting attribute ranking methods that help determine the relative importance of rule antecedent attributes involved in a sparse rule base. The approach has been developed to form a generalised methodology from two-fold perspectives: 1) it allows for any established ranking method to be utilised to score the attributes, leading to a flexible weighting scheme for FRI; and 2) it can be extended to any other FRI which involves multiple rule antecedents but not assigned with individual weights.

The implemented work also enables fuzzy rule based systems to use the conventional compositional rule of inference (CRI) and the weighted FRI jointly. Through this integration an implemented system can obtain more accurate inference results, thereby taking advantage of both methods: CRI for matched observations and weighted FRI for unmatched ones. The following summarises the main work from this research, reflecting its key contributions for achieving the goal of attribute weighted fuzzy interpolative reasoning. In particular, the summarising discussions are presented in response to the original two research aims as identified in Section 1.3: 1) how the weights are generated; and 2) whether and how the weights are integrated within the underlying, non-weighted FRI.

8.1.1 Generation of Attribute Weights

In response to this issue, without requiring any observation or running the underlying FRI system, the proposed weight learning method can automatically determine the relative importance of rule antecedent attributes by the use of the given sparse rule base only. An innovative reverse engineering procedure has been proposed, through which to compute the ranking scores from an artificial decision table derived from the given rules. Such a learning method is independent from the underlying FRI mechanism, thereby offering flexibility in developing fuzzy systems. To reflect this viewpoint, the thesis has provided several different attribute ranking methods as alternatives for attribute weighting, based on popular feature selection techniques in the relevant literature.

8.1.2 Integration of Weights with Fuzzy Interpolative Reasoning

In dealing with the second identified challenging problem (of whether and how weights are integrated within an FRI that works effectively), three pieces of distinctive work have been carried out in this research:

- Weighting Fuzzy Rule Interpolation

Given the generated weights of rule antecedent attributes, the scale and move transformation-based FRI (T-FRI) has been first adopted as an initial investigation to develop a weighted FRI algorithm, with the aforementioned weighting mechanism thoroughly applied within each core step of T-FRI. The weighted fuzzy interpolative reasoning has also been established by extending weighted T-FRI to two other classical FRI methods, namely the KH and CCL algorithms. Together, these weighted FRI approaches provide possible alternatives for implementing the proposed weighted fuzzy rule-based interpolative framework that works effectively.

- Interpolating with Just Two Nearest Neighbouring Weighted Fuzzy Rules

The proposed weighted FRI algorithms, namely the weighted T-FRI, weighted KH and weighted CCL, have been systematically evaluated through experimentation. The results demonstrate the superior performance of these weighted methods over their original unweighted counterparts, without incurring significant increase in run time cost. These collectively reflect the effectiveness and efficiency of weighted FRI. Very importantly, as have been illustrated by experimental analysis, supported by attribute ranking, only two (i.e., the least number of) closest rules are required to perform accurate interpolation. As such, better results can be achieved with fewest rules. This helps increase computational efficiency, without the need of searching for and operating on multiple rules beyond the immediate neighborhood of a given observation. It also helps reinforce the stability of the underlying weighted FRI mechanism.

- Applying Weighted Fuzzy Interpolative Reasoning

The proposed weighted fuzzy interpolative reasoning framework has been fully implemented (using different feature evaluation and unweighted FRI techniques) and successfully applied for classification and prediction problems. It has been systematically shown to outperform both unweighted FRI and the state-of-the-art weighted FRI techniques. Furthermore, a novel fuzzy rule-based system for interpretable mammographic mass classification has been presented. This provides another case to demonstrate the potential success in applying weighted fuzzy interpolative reasoning in practical problem settings.

8.2 Future Work

Whilst very promising, there is much room to strengthen the work reported in this thesis. This final section prints out important further work that of carried out, will improve the present research.

8.2.1 Weights Generation via Reverse Engineering

This thesis assumes that in general, a sparse rule base is given for producing the weights of rule antecedent attributes. Yet, in the experimental evaluation for the proposed work, a data-driven rule learning mechanism is presumed available to convert a given dataset into rules, with a simple fuzzification procedure. Nevertheless, the attribute weights are learned by the use of the rule base only. Thus, it would be interesting to investigate how much better a weighted FRI method may perform with optimized fuzzy quantities and rules.

Additionally, the problem of the curse of dimensionality may arise due to the production of the artificial training instances from the given rule base, as the number of missing rule antecedents increases despite only a sparse rule base is involved. Thus, it is desirable to increase the algorithmic efficiency while revising the work. Potential solutions to this include: to exploit feature selection techniques (e.g., [Jensen and Shen, 2009, Diao and Shen, 2015]) to restrain the learning process; and to explore link-based analysis tools (e.g., [Boongoen et al., 2010, Shen and Boongoen, 2012]) to better associate and refine the rules and rule conditions.

8.2.2 Weighted Fuzzy Interpolative Reasoning

The proposed methodology for weighted FRI can be further enhanced from the following viewpoints.

8.2.2.1 Weighting Dynamic FRI

The proposed weighted transformation-based FRI currently works on a static rule base. Yet, a volume of intermediate fuzzy rules are typically generated while executing rule interpolation. From this, the ideas of [Naik et al., 2017b] can be exploited to enrich the rule base by refining and promoting these intermediate rules, gaining efficiency by allowing for more direct rule-firing without running the interpolation procedure. In particular, the attribute weights in the present work may help leading to a weighted assembly of additional rules, thereby improving the performance of the emerged rule base by considering different importance levels amongst the rule antecedents. Nonetheless, in general, any addition or removal of certain original rules will affect the weights induced from the given rule base, which in turn, will affect the interpolated results. The exact influence upon the interpolative reasoning process therefore, remains a piece of important further research.

8.2.2.2 Theoretical Analysis of "Two Rules Interpolation"

The conjecture that "least number of neighboring rules do better with weighted FRI" has been empirically shown to hold for three weighted FRI approaches. Whilst this has been supported by substantial and consistent experimental results, it is unclear how to further verify this notion through mathematically rigorous analysis. This forms an important next step to reinforce the current research.

8.2.2.3 Use of Non-triangular Fuzzy Sets

As stated throughout this thesis, all implementations of the proposed approaches have been carried out on the basis of representing fuzzy values with triangular membership functions. It has a natural appeal to consider modifying the implemented systems with more sophisticated and more powerful representations of fuzzy values, which would not too much additional computation overheads. Trapezoidal and Gaussian representations are likely candidates for this. An investigation into how they be utilised to enable the desirable improvement forms another piece of further research.

8.2.2.4 Weighted FRI with TSK Fuzzy Models

All work carried out so far has to do with fuzzy rules of Mamdani type. Most recently, there has been research which reports on extending conventional T-FRI methods to building FRI mechanisms for Takagi Sugeno Kang (TSK) fuzzy models in general [Chen et al., 2019] and ANFIS (Adaptive-Network-based Fuzzy Inference System) in particular [Yang et al., 2018]. The extension of rule interpolation on ANFIS has also seen a successful initial application for addressing image super resolution problem [Yang et al., 2019]. Nevertheless, all of these developments follow an unweighted approach. Thus, it would be very interesting to consider further extending such work within the weighted FRI framework.

Appendix A

Iterative Rule Base Generation

A data-driven rule base learning mechanism is to generate rules by generalising raw data, with rules expressed in the format of antecedents associated with a corresponding consequent [Hong and Lee, 1996, Wang and Mendel, 1992]. Such a generation process may follow an iterative procedure [Galea and Shen, 2006, Hoffmann, 2004] to incrementally add new rules to the rule base. This appendix outlines an iterative rule base generation procedure, which repeatedly sequentially extracts rules from data into an emerging rule base, and which is utilised in this thesis for producing rules in all experimental studies (unless otherwise stated).

Given a set of instances each of which consists of r antecedent attributes and a consequent attribute, a rule base is generated in an iterative procedure as illustrated in Algorithm 10. Here, fuzzy rules are considered for generality, which may be readily degenerated into a crisp rule set if preferred. The iteration process is terminated by checking against a pre-set threshold value, that determines at least how many data points have been covered by the extracted rules so far.

Before the iterative procedure is executed to generate the rule base, the domains of all r antecedent attributes and the consequent attribute are quantified evenly into $m_1, m_2, \ldots, m_r$ and m_c fuzzy regions, respectively, where m_c denotes the number of regions for the consequent attribute. Each fuzzy region is assigned with a membership function (implemented with triangular membership functions in this work for consistency and simplicity). This results in a division of fuzzy region space of the antecedent of an emerging rule in the form of a hypercube, of which each hypergrid stands for a combination of particular fuzzy regions of the r antecedent attributes.

Algorithm 10 Iterative Rule Induction from Data

Input:

- Data set of instances D
- Threshold value δ

Output:

- Rule base R

Divide the domain of each antecedent and consequent attribute evenly into a certain number of fuzzy regions, and construct the fuzzy region space (FRS) of the antecedent, which is a hypercube with the dimensionality of $m1 \times m2 \cdots \times m_r$, where $m_i, i = 1, \ldots, r$ stands for the number of regions for the ith attribute;
while true **do**
  Apply the data of instances D into the FRS and match each instance to a corresponding hypergrid in the hypercube in terms of its antecedent attributes;
  Select the hypergrid with the highest hits, denoting n as the highest hit number;
  if $n > \delta$ **then**
    Extract a rule from this hypergrid, and add it into the rule base R;
    Remove all of the instances which hit this hypergrid from D, update $D = D - D_{most\_hit}$;
  else
    End While Loop;
  end if
end while
Return R

The iteration process begins with a complete data set of instances D. A hypergrid hit by an instance indicates a certain value of membership is obtained for the corresponding combination of fuzzy regions. The hypergrid which is most covered by the instances in D receives the most hits amongst all. As indicated above, a threshold δ is used to determine whether the most covered hypergrid can form a rule and be added into the emerging rule base R. If the number of the highest hits is larger than the threshold, a rule is extracted from this hypergrid.

The rule antecedent values returned by one iteration are those fuzzy values associated with the corresponding hypergrid. The rule consequent adopts the fuzzy value which corresponds to one of the m_c values at which the instances have the highest number of hits. After this, those instances hit in this hypergrid are removed from the original data set, and the iterative process repeats by treating the remaining data as the input data set to start the next round for the generation of the rules following the current one. However, if the proportion of hit instances is less than δ, a rule cannot be generated by this hypergrid because a small number of those hits may just be due to noise, and the iterative procedure is hence continued to the next round until all given instances are removed.

As stated earlier, this simple iterative rule generation procedure is used to learn a rule base to construct the inference system proposed in this thesis, especially for the implementations which assumed that no rules were provided by domain experts (see Chapter 5). Of course, a number of other advanced rule learning methodologies can be adopted to generate more compact rule bases, but this is beyond the scope of this thesis.

Appendix B

List of Acronyms

ANFIS Adaptive Neuro-Fuzzy Inference System

AUC Area Under Curve

AW Attribute Weight

BCDR Breast Cancer Digital Repository

BI-RADS Breast Imaging Reporting and Data System

CADx Computer-Aided Diagnosis

CC S.-M. Chen and Z.-J. Chen

CCL Chang, Chen and Liau

CFS Correlation-based Feature Selection

COG Center Of Gravity

CP(s) Characteristic Point(s)

CRI Compositional Rule of Inference

CV Cross Validation

DCNNs Deep Convolutional Neural Networks

DM Digital Mammography

FCM Fuzzy C-Means

FM Film Mammography

FN False Negative

FP False Positive

FRFS Fuzzy-Rough Feature Selection

FRI Fuzzy Rule Interpolation

FS Feature Selection

GA Genetic Algorithm

IG Information Gain

IG-T-FRI Information Gain-guided T-FRI

IR Inconsistency Rate

IRFS Consistency-based Feature Selection

KEEL Knowledge Extraction based on Evolutionary Learning

KH Kóczy and Hirota

LLC Local Learning based Clustering

LLCFS Local Learning-based Clustering for Feature Selection

LS Laplacian Score

MF Membership Function

MMC Mammographic Mass Classification

NRL Normalised Radial Length

Rep Representative Value

RF Ratio of Fuzziness

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic

ROI Region Of Interest

RSFS Rough Set-based Feature Selection

SD Standard Deviation

SRM Semantic Revision Method

T-FRI Scale and Move Transformation-based Fuzzy Rule Interpolation

TN True Negative

TP True Positive

Bibliography

[Afify, 2016] Afify, A. A. (2016). A fuzzy rule induction algorithm for discovering classification rules. *Journal of Intelligent & Fuzzy Systems*, 30(6):3067–3085.

[Ahmad et al., 2016] Ahmad, B., Murphy, J., Langdon, P., and Godsill, S. (2016). Bayesian intent prediction in object tracking using bridging distributions. *IEEE transactions on cybernetics*.

[Al-Sharhan et al., 2001] Al-Sharhan, S., Karray, F., Gueaieb, W., and Basir, O. (2001). Fuzzy entropy: a brief survey. In *10th IEEE International Conference on Fuzzy Systems.(Cat. No. 01CH37297)*, volume 3, pages 1135–1139. IEEE.

[Alcalá-Fdez et al., 2011] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.

[Baranyi et al., 1995] Baranyi, P., Gedeon, T., and Kóczy, L. (1995). A general method for fuzzy rule interpolation: specialized for crisp triangular and trapezoidal rules. In *EUFIT'95*, pages 99–102.

[Baranyi et al., 1996a] Baranyi, P., Gedeon, T., and Koczy, L. (1996a). Rule interpolation by spatial geometric representation. In *Proc. IPMU*, volume 96, pages 483–488.

[Baranyi et al., 1996b] Baranyi, P., Gedeon, T. D., and Kóczy, L. T. (1996b). A general interpolation technique in fuzzy rule bases with arbitrary membership functions. In *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No. 96CH35929)*, volume 1, pages 510–515. IEEE.

[Baranyi and Kóczy, 1996a] Baranyi, P. and Kóczy, L. (1996a). Multidimensional fuzzy rule inter-and extrapolation based on geometric solution. In *Proc. 7th Int. Power Electronics and Motion Control Conference (PEMC'96)*, volume 3, pages 443–447.

[Baranyi et al., 2004] Baranyi, P., Kóczy, L., and Gedeon, T. (2004). A generalized concept for fuzzy rule interpolation. *IEEE Transactions on Fuzzy Systems*, 12(6):820–837.

[Baranyi et al., 1998] Baranyi, P., Mizik, S., Kóczy, L. T., Gedeon, T. D., and Nagy, I. (1998). Fuzzy rule base interpolation based on semantic revision. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, volume 2, pages 1306–1311. IEEE.

[Baranyi and Kóczy, 1996b] Baranyi, P. Z. and Kóczy, L. (1996b). A general and specialised solid cutting method for fuzzy rule interpolation. *Journal BUSEFAL, URA-CNRS, Université Paul Sabatier*, pages 13–22.

[Baranyi et al., 1999] Baranyi, P. Z., Tikk, D., Yam, Y., Kóczy, L., and Nádai, L. (1999). A new method for avoiding abnormal conclusion for a-cut based rule interpolation. In *1999 IEEE international fuzzy systems conference, FUZZ-IEEE '99. Seoul*, pages 1383–1388. IEEE.

[Bede, 2013] Bede, B. (2013). *Mathematics of Fuzzy Sets and Fuzzy Logic*. Springer-Verlag Berlin Heidelberg.

[Berecz, 2009] Berecz, A. (2009). Fuzzy rule interpolation based tool life modeling using rbesi and fripoc. In *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, pages 11–16. IEEE.

[Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203.

[Bonissone, 1997] Bonissone, P. P. (1997). Soft computing: the convergence of emerging reasoning technologies. *Soft computing*, 1(1):6–18.

[Boongoen et al., 2010] Boongoen, T., Shen, Q., and Price, C. (2010). Disclosing false identity through hybrid link analysis. *Artificial Intelligence and Law*, 18(1):77–102.

[Bouchon-Meunier et al., 1999] Bouchon-Meunier, B., Delechamp, J., Marsala, C., Mellouli, N., Rifqi, M., and Zerrouki, L. (1999). Analogy and fuzzy interpolation in the case of sparse rules. In *Proc. Joint Conf. Eurofuse-Soft and Intelligent Computing*, pages 132–136.

[Bouchon-Meunier et al., 2001] Bouchon-Meunier, B., Dubois, D., Marsala, C., Prade, H., and Ughetto, L. (2001). A comparative view of interpolation methods between sparse fuzzy rules. In *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, pages 2499–2504. IEEE.

[Bouchon-Meunier et al., 2000] Bouchon-Meunier, B., Marsala, C., and Rifqi, M. (2000). Interpolative reasoning based on graduality. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 1, pages 483–487. IEEE.

[Bouchon-Meunier and Valverde, 1999] Bouchon-Meunier, B. and Valverde, L. (1999). A fuzzy approach to analogical reasoning. *Soft Computing*, 3(3):141–147.

[Box et al., 2015] Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[Burkhardt and Bonissone, 1992] Burkhardt, D. G. and Bonissone, P. P. (1992). Automated fuzzy knowledge base generation and tuning. In *[1992 Proceedings] IEEE International Conference on Fuzzy Systems*, pages 179–188. IEEE.

[Chang and Chen, 2008] Chang, Y.-C. and Chen, S.-M. (2008). A new method for multiple fuzzy rules interpolation with weighted antecedent variables. In *2008 IEEE International Conference on Systems, Man and Cybernetics*, pages 76–81. IEEE.

[Chang et al., 2008] Chang, Y.-C., Chen, S.-M., and Liau, C.-J. (2008). Fuzzy interpolative reasoning for sparse fuzzy-rule-based systems based on the areas of fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 16(5):1285–1301.

[Chapra and Canale, 1998] Chapra, S. C. and Canale, R. P. (1998). *Numerical methods for engineers*, volume 2. Mcgraw-hill New York.

[Chen et al., 2016] Chen, C., Parthaláin, N. M., Li, Y., Price, C., Quek, C., and Shen, Q. (2016). Rough-fuzzy rule interpolation. *Information Sciences*, 351:1–17.

[Chen and Shen, 2012] Chen, C. and Shen, Q. (2012). A new method for rule interpolation inspired by rough-fuzzy sets. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–8. IEEE.

[Chen and Shen, 2017] Chen, C. and Shen, Q. (2017). Transformation-based fuzzy rule interpolation using interval type-2 fuzzy sets. *Algorithms*, 10(3):91, 20pp.

[Chen and Wang, 1999] Chen, M.-S. and Wang, S.-W. (1999). Fuzzy clustering analysis for optimizing fuzzy membership functions. *Fuzzy sets and systems*, 103(2):239–254.

[Chen and Adam, 2017] Chen, S.-M. and Adam, S. I. (2017). Adaptive fuzzy interpolation based on general representative values of polygonal fuzzy sets and the shift and modification techniques. *Information Sciences*, 414:147–157.

[Chen and Adam, 2018] Chen, S.-M. and Adam, S. I. (2018). Weighted fuzzy interpolated reasoning based on ranking values of polygonal fuzzy sets and new scale and move transformation techniques. *Information Sciences*, 435:184–202.

[Chen and Chang, 2011a] Chen, S.-M. and Chang, Y.-C. (2011a). Weighted fuzzy interpolative reasoning for sparse fuzzy rule-based systems. *Expert Systems with Applications*, 38(8):9564–9572.

[Chen and Chang, 2011b] Chen, S.-M. and Chang, Y.-C. (2011b). Weighted fuzzy rule interpolation based on ga-based weight-learning techniques. *IEEE Transactions on Fuzzy Systems*, 19(4):729–744.

[Chen et al., 2013a] Chen, S.-M., Chang, Y.-C., Chen, Z.-J., and Chen, C.-L. (2013a). Multiple fuzzy rules interpolation with weighted antecedent variables in sparse fuzzy rule-based systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(05):1359002.

[Chen and Chen, 2016] Chen, S.-M. and Chen, Z.-J. (2016). Weighted fuzzy interpolative reasoning for sparse fuzzy rule-based systems based on piecewise fuzzy entropies of fuzzy sets. *Information Sciences*, 329:503–523.

[Chen et al., 2015] Chen, S.-M., Cheng, S.-H., and Chen, Z.-J. (2015). Fuzzy interpolative reasoning based on the ratio of fuzziness of rough-fuzzy sets. *Information Sciences*, 299:394–411.

[Chen and Hsin, 2014] Chen, S.-M. and Hsin, W.-C. (2014). Weighted fuzzy interpolative reasoning based on the slopes of fuzzy sets and particle swarm optimization techniques. *IEEE transactions on cybernetics*, 45(7):1250–1261.

[Chen and Ko, 2008] Chen, S.-M. and Ko, Y.-K. (2008). Fuzzy interpolative reasoning for sparse fuzzy rule-based systems based on α-cuts and transformations techniques. *IEEE Transactions on Fuzzy Systems*, 16(6):1626–1648.

[Chen et al., 2009] Chen, S.-M., Ko, Y.-K., Chang, Y.-C., and Pan, J.-S. (2009). Weighted fuzzy interpolative reasoning based on weighted increment transformation and weighted ratio transformation techniques. *IEEE Transactions on Fuzzy Systems*, 17(6):1412–1427.

[Chen and Lee, 2011] Chen, S.-M. and Lee, L.-W. (2011). Fuzzy interpolative reasoning for sparse fuzzy rule-based systems based on interval type-2 fuzzy sets. *Expert Systems with Applications*, 38(8):9947–9957.

[Chen et al., 2013b] Chen, S.-M., Lee, L.-W., and Shen, V. R. (2013b). Weighted fuzzy interpolative reasoning systems based on interval type-2 fuzzy sets. *Information Sciences*, 248:15–30.

[Chen et al., 2019] Chen, T., Shang, C., Yang, J., Li, F., and Shen, Q. (2019). A new approach for transformation-based fuzzy rule interpolation. *IEEE Transactions on Fuzzy Systems*.

[Cheng et al., 2006] Cheng, H.-D., Shi, X., Min, R., Hu, L., Cai, X., and Du, H. (2006). Approaches for automated detection and classification of masses in mammograms. *Pattern recognition*, 39(4):646–668.

[Cheng et al., 2015] Cheng, S.-H., Chen, S.-M., and Chen, C.-L. (2015). Fuzzy interpolative reasoning based on ranking values of polygonal fuzzy sets and automatically generated weights of fuzzy rules. *Information Sciences*, 325:521–540.

[Cheng et al., 2016] Cheng, S.-H., Chen, S.-M., and Chen, C.-L. (2016). Adaptive fuzzy interpolation based on ranking values of polygonal fuzzy sets and similarity measures between polygonal fuzzy sets. *Information Sciences*, 342:176–190.

[Chouchoulas and Shen, 2001] Chouchoulas, A. and Shen, Q. (2001). Rough set-aided keyword reduction for text categorization. *Applied Artificial Intelligence*, 15(9):843–873.

[CROWDER, 1990] CROWDER, R. (1990). Predicting the mackey-glass time series with cascade correlation learning. In *Proceedings 1990 Connectionist Models Summer School*, pages 117–123. Morgan Kaufmann.

[Cui et al., 2010] Cui, Y., Jin, J. S., Zhang, S., Luo, S., and Tian, Q. (2010). Correlation-based feature selection and regression. In *Pacific-Rim Conference on Multimedia*, pages 25–35. Springer.

[Dash and Liu, 1997] Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156.

[Dash and Liu, 2003] Dash, M. and Liu, H. (2003). Consistency-based search in feature selection. *Artificial intelligence*, 151(1-2):155–176.

[De Kleer, 1986] De Kleer, J. (1986). An assumption-based tms. *Artificial intelligence*, 28(2):127–162.

[De Luca and Termini, 1972] De Luca, A. and Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and control*, 20(4):301–312.

[Dheeru and Karrai, 2017] Dheeru, D. and Karrai, T. (2017). UCI machine learning repository.

[Diao et al., 2014] Diao, R., Jin, S., and Shen, Q. (2014). Antecedent selection in fuzzy rule interpolation using feature selection techniques. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 2206–2213. IEEE.

[Diao and Shen, 2015] Diao, R. and Shen, Q. (2015). Nature inspired feature selection meta-heuristics. *Artificial Intelligence Review*, 44(3):311–340.

[Ding et al., 1989] Ding, L., Shen, Z., and Mukaidono, M. (1989). A new method for approximate reasoning. In *Proceedings. The Nineteenth International Symposium on Multiple-Valued Logic*, pages 179–185. IEEE.

[Ding et al., 1992] Ding, L., Shen, Z., and Mukaidono, M. (1992). Revision principle for approximate reasoning, based on linear revising method. In *Proceedings of the Second International Conference on Fuzzy Logic and Neural Networks, Iizuka'92*, pages 305–308.

[Dominguez and Nandi, 2008] Dominguez, A. R. and Nandi, A. K. (2008). Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection. *Computerized Medical Imaging and Graphics*, 32(4):304–315.

[Eshelman, 1991] Eshelman, L. J. (1991). The chc adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. In *Foundations of genetic algorithms*, volume 1, pages 265–283. Elsevier.

[Fu and Shen, 2010] Fu, X. and Shen, Q. (2010). Fuzzy compositional modeling. *IEEE Transactions on Fuzzy Systems*, 18(4):823–840.

[Fuller, 2000] Fuller, R. (2000). *Introduction to neuro-fuzzy systems*, volume 2. Springer Science & Business Media.

[Galea and Shen, 2006] Galea, M. and Shen, Q. (2006). Simultaneous ant colony optimization algorithms for learning linguistic fuzzy rules. *Swarm intelligence in data mining*, pages 75–99.

[Ghiselli, 1964] Ghiselli, E. (1964). Theory of psychological measurement.

[Hall, 2000] Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

[He et al., 2006] He, X., Cai, D., and Niyogi, P. (2006). Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514.

[Hoffmann, 2004] Hoffmann, F. (2004). Combining boosting and evolutionary algorithms for learning of fuzzy classification rules. *Fuzzy Sets and Systems*, 141(1):47–58.

[Holland, 1975] Holland, J. (1975). Adaptation in natural and artificial systems. *Holland, JH*.

[Hong and Lee, 1996] Hong, T.-P. and Lee, C.-Y. (1996). Induction of fuzzy rules and membership functions from training examples. *Fuzzy sets and Systems*, 84(1):33–47.

[Hsiao et al., 1998] Hsiao, W.-H., Chen, S.-M., and Lee, C.-H. (1998). A new interpolative reasoning method in sparse rule-based systems. *Fuzzy Sets and Systems*, 93(1):17–22.

[Huang et al., 2004] Huang, D.-M., Tsang, E., and Yeung, D. (2004). A fuzzy interpolative reasoning method. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*, volume 3, pages 1826–1830. IEEE.

[Huang and Shen, 2003] Huang, Z. and Shen, Q. (2003). A new fuzzy interpolative reasoning method based on center of gravity. In *The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03.*, volume 1, pages 25–30. IEEE.

[Huang and Shen, 2006] Huang, Z. and Shen, Q. (2006). Fuzzy interpolative reasoning via scale and move transformations. *IEEE Transactions on Fuzzy Systems*, 14(2):340–359.

[Huang and Shen, 2008] Huang, Z. and Shen, Q. (2008). Fuzzy interpolation and extrapolation: A practical approach. *IEEE Transactions on Fuzzy Systems*, 16(1):13–28.

[Janikow, 1998] Janikow, C. Z. (1998). Fuzzy decision trees: issues and methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(1):1–14.

[Jenei, 2001] Jenei, S. (2001). Interpolation and extrapolation of fuzzy quantities revisited–an axiomatic approach. *Soft Computing*, 5(3):179–193.

[Jenei et al., 2002] Jenei, S., Klement, E.-P., and Konzel, R. (2002). Interpolation and extrapolation of fuzzy quantities–the multiple-dimensional case. *Soft Computing*, 6(3-4):258–270.

[Jensen and Shen, 2004] Jensen, R. and Shen, Q. (2004). Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. *Knowledge and Data Engineering, IEEE Transactions on*, 16(12):1457–1471.

[Jensen and Shen, 2007] Jensen, R. and Shen, Q. (2007). Fuzzy-rough sets assisted attribute selection. *Fuzzy Systems, IEEE Transactions on*, 15(1):73–89.

[Jensen and Shen, 2008] Jensen, R. and Shen, Q. (2008). *Computational intelligence and feature selection: rough and fuzzy approaches*, volume 8. John Wiley & Sons.

[Jensen and Shen, 2009] Jensen, R. and Shen, Q. (2009). New approaches to fuzzy-rough feature selection. *Fuzzy Systems, IEEE Transactions on*, 17(4):824–838.

[Jin et al., 2014] Jin, S., Diao, R., Quek, C., and Shen, Q. (2014). Backward fuzzy rule interpolation. *IEEE Transactions on Fuzzy Systems*, 22(6):1682–1698.

[Jin et al., 2020] Jin, S., Peng, J., Li, Z., and Shen, Q. (2020). Backward fuzzy rule interpolation. *Information Sciences*, 506:99–112.

[Jin et al., 2019] Jin, S., Shen, Q., and Peng, J. (2019). Transformation based backward fuzzy rule interpolation with multiple missing antecedent values. In *Backward Fuzzy Rule Interpolation*, pages 75–89. Springer.

[Johanyák, 2010] Johanyák, Z. C. (2010). Student evaluation based on fuzzy rule interpolation. *Int. J. Artif. Intell.*, 5(A10):37–55.

[Johanyák and Kovács, 2006] Johanyák, Z. C. and Kovács, S. (2006). Survey on various interpolation based fuzzy reasoning methods. *Production Systems and Information Engineering*, 3(1):39–56.

[Kawaguchi and Miyakoshi, 2000a] Kawaguchi, M. F. and Miyakoshi, M. (2000a). A fuzzy rule interpolation technique based on bi-splines in multiple input systems. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 1, pages 488–492. IEEE.

[Kawaguchi and Miyakoshi, 2000b] Kawaguchi, M. F. and Miyakoshi, M. (2000b). Fuzzy spline functions and nonlinear interpolation of if-then rules. In *Proc. Int. Conf. Soft Computing*, pages 785–792.

[Kawaguchi and Miyakoshi, 2001] Kawaguchi, M. F. and Miyakoshi, M. (2001). Fuzzy spline interpolation in sparse fuzzy rule bases. In *A New Paradigm Of Knowledge Engineering By Soft Computing*, pages 95–120. World Scientific.

[Kawaguchi et al., 1997] Kawaguchi, M. F., Miyakoshi, M., and Kawaguchi, M. (1997). Linear interpolation with triangular rules in sparse fuzzy rule bases. *Proc. IFSA World Congr*, pages 138–143.

[Kira and Rendell, 1992a] Kira, K. and Rendell, L. A. (1992a). The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134.

[Kira and Rendell, 1992b] Kira, K. and Rendell, L. A. (1992b). A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pages 249–256. Elsevier.

[Kóczy and Hirota, 1991] Kóczy, L. and Hirota, K. (1991). Rule interpolation by α-level sets in fuzzy approximate reasoning. *J. BUSEFAL, Automne, URA-CNRS*, 46:115–123.

[Kóczy and Hirota, 1993a] Kóczy, L. and Hirota, K. (1993a). Approximate reasoning by linear rule interpolation and general approximation. *International Journal of Approximate Reasoning*, 9(3):197–225.

[Kóczy and Hirota, 1993b] Kóczy, L. and Hirota, K. (1993b). Interpolative reasoning with insufficient evidence in sparse fuzzy rule bases. *Information Sciences*, 71(1):169–201.

[Kóczy and Hirota, 1993c] Kóczy, L. and Hirota, K. (1993c). Ordering, distance and closeness of fuzzy sets. *Fuzzy sets and systems*, 59(3):281–293.

[Koczy and Hirota, 1997] Koczy, L. T. and Hirota, K. (1997). Size reduction by interpolation in fuzzy rule bases. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(1):14–25.

[Kóczy et al., 2000] Kóczy, L. T., Hirota, K., and Gedeon, T. D. (2000). Fuzzy rule interpolation by the conservation of relative fuzziness. *JACIII*, 4(1):95–101.

[Koczy et al., 2000] Koczy, L. T., Hirota, K., and Muresan, L. (2000). Interpolation in hierarchical fuzzy rule bases. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 1, pages 471–477. IEEE.

[Kononenko, 1994] Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer.

[Kovács, 2006] Kovács, S. (2006). Extending the fuzzy rule interpolation" five" by fuzzy observation. In *Computational Intelligence, Theory and Applications*, pages 485–497. Springer.

[Kovács and Kóczy, 1997] Kovács, S. and Kóczy, L. T. (1997). Approximate fuzzy reasoning based on interpolation in the vague environment of the fuzzy rule base as a practical alternative of the classical cri. In *Proceedings of the 7th International Fuzzy Systems Association World Congress, Prague, Czech Republic*, pages 144–149.

[Kovacs and Koczy, 1997] Kovacs, S. and Koczy, L. T. (1997). Approximate fuzzy reasoning based on interpolation in the vague environment of the fuzzy rulebase. In *Proceedings of IEEE International Conference on Intelligent Engineering Systems*, pages 63–68. IEEE.

[Kovács and Kóczy, 1997] Kovács, S. and Kóczy, L. T. (1997). The use of the concept of vague environment in approximate fuzzy reasoning. *Fuzzy Set Theory and Applications, Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Bratislava, Slovak Republic*, 12:169–181.

[Kovács and Kóczy, 1999] Kovács, S. and Kóczy, L. T. (1999). Application of an approximate fuzzy logic controller in an agv steering system, path tracking and collision avoidance strategy. *Fuzzy Set Theory and Applications, In Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences*, 16:456–467.

[Kovács and Kóczy, 2004] Kovács, S. and Kóczy, L. T. (2004). Application of interpolation-based fuzzy logic reasoning in behaviour-based control structures. In *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542)*, volume 3, pages 1543–1548. IEEE.

[Lévy and Jain, 2016] Lévy, D. and Jain, A. (2016). Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*.

[Li et al., 2018a] Li, F., Li, Y., Shang, C., and Shen, Q. (2018a). Improving fuzzy rule interpolation performance with information gain-guided antecedent weighting. *Soft Computing*, 22(10):3125–3139.

[Li et al., 2019a] Li, F., Li, Y., Shang, C., and Shen, Q. (2019a). Fuzzy knowledge-based prediction through weighted rule interpolation. *IEEE Transactions on Cybernetics*.

[Li et al., 2017a] Li, F., Shang, C., Li, Y., and Shen, Q. (2017a). Feature ranking-guided fuzzy rule interpolation. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.

[Li et al., 2017b] Li, F., Shang, C., Li, Y., and Shen, Q. (2017b). Guiding fuzzy rule interpolation with information gains. In *Advances in Computational Intelligence Systems*, pages 165–183. Springer.

[Li et al., 2018b] Li, F., Shang, C., Li, Y., and Shen, Q. (2018b). Feature ranking-guided fuzzy rule interpolation for mammographic mass shape classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE.

[Li et al., 2019b] Li, F., Shang, C., Li, Y., and Shen, Q. (2019b). Interpretable mammographic mass classification with fuzzy interpolative reasoning. *Knowledge-Based Systems*, page 105279.

[Li et al., 2018c] Li, F., Shang, C., Li, Y., Yang, J., and Shen, Q. (2018c). Fuzzy rule based interpolative reasoning supported by attribute ranking. *IEEE Transactions on Fuzzy Systems*, 26(5):2758–2773.

[Li et al., 2019c] Li, F., Shang, C., Li, Y., Yang, J., and Shen, Q. (2019c). Interpolation with just two nearest neighbouring weighted fuzzy rules. *IEEE Transactions on Fuzzy Systems*.

[Li et al., 2016] Li, F., Shen, Q., Li, Y., and Mac Parthaláin, N. (2016). Handwritten chinese character recognition using fuzzy image alignment. *Soft Computing*, 20(8):2939–2949.

[Li et al., 2005] Li, Y.-M., Huang, D.-M., and Zhang, L.-N. (2005). Weighted fuzzy interpolative reasoning method. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 5, pages 3104–3108. IEEE.

[Liu and Motoda, 2012] Liu, H. and Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media.

[Liu and Tang, 2013] Liu, X. and Tang, J. (2013). Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. *IEEE Systems Journal*, 8(3):910–920.

[Lopez et al., 2012] Lopez, M. G., Posada, N., Moura, D. C., Pollán, R. R., Valiente, J. M. F., Ortega, C. S., Solar, M., Diaz-Herrero, G., Ramos, I., Loureiro, J., et al. (2012). Bcdr: a breast cancer digital repository. In *15th International conference on experimental mechanics*.

[Magna et al., 2016] Magna, G., Casti, P., Jayaraman, S. V., Salmeri, M., Mencattini, A., Martinelli, E., and Di Natale, C. (2016). Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system. *Knowledge-Based Systems*, 101:60–70.

[Mamdani and Assilian, 1999] Mamdani, E. and Assilian, S. (1999). An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of human-computer studies*, 51(2):135–147.

[Mendel et al., 2006] Mendel, J. M., John, R. I., and Liu, F. (2006). Interval type-2 fuzzy logic systems made simple. *IEEE transactions on fuzzy systems*, 14(6):808–821.

[Miranda and Felipe, 2015] Miranda, G. H. B. and Felipe, J. C. (2015). Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization. *Computers in biology and medicine*, 64:334–346.

[Mitchell, 1997] Mitchell, T. M. (1997). Machine learning.

[Moura and López, 2013] Moura, D. C. and López, M. A. G. (2013). An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International journal of computer assisted radiology and surgery*, 8(4):561–574.

[Moura et al., 2013] Moura, D. C., López, M. A. G., Cunha, P., de Posada, N. G., Pollan, R. R., Ramos, I., Loureiro, J. P., Moreira, I. C., de Araújo, B. M. F., and Fernandes, T. C. (2013). Benchmarking datasets for breast cancer computer-aided diagnosis (cadx). In *Iberoamerican Congress on Pattern Recognition*, pages 326–333. Springer.

[Mousa et al., 2005] Mousa, R., Munib, Q., and Moussa, A. (2005). Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. *Expert systems with Applications*, 28(4):713–723.

[Mu et al., 2008] Mu, T., Nandi, A. K., and Rangayyan, R. M. (2008). Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers. *Journal of Digital Imaging*, 21(2):153–169.

[Mukaidono et al., 1990] Mukaidono, M., Ding, L., and Shen, Z. (1990). Approximate reasoning based on revision principle. In *Proc. NAFIPS'90*, volume 1, pages 94–97.

[Nagy et al., 2018] Nagy, S., Sziová, B., and Kóczy, L. T. (2018). The effect of image feature qualifiers on fuzzy colorectal polyp detection schemes using kh interpolation-towards hierarchical fuzzy classification of coloscopic still images. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE.

[Naik et al., 2017a] Naik, N., Diao, R., Shang, C., Shen, Q., and Jenkins, P. (2017a). D-fri-winfirewall: Dynamic fuzzy rule interpolation for windows firewall. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.

[Naik et al., 2014] Naik, N., Diao, R., and Shen, Q. (2014). Genetic algorithm-aided dynamic fuzzy rule interpolation. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2198–2205. IEEE.

[Naik et al., 2017b] Naik, N., Diao, R., and Shen, Q. (2017b). Dynamic fuzzy rule interpolation and its application to intrusion detection. *IEEE Transactions on Fuzzy Systems*, 26(4):1878–1892.

[Oliver et al., 2010] Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J., R.E. Denton, E., and Zwiggelaar, R. (2010). A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis*, 14(2):87–110.

[Oliver et al., 2012] Oliver, A., Torrent, A., Lladó, X., Tortajada, M., Tortajada, L., Sentís, M., Freixenet, J., and Zwiggelaar, R. (2012). Automatic microcalcification and cluster detection for digital and digitised mammograms. *Knowledge-Based Systems*, 28:68–75.

[Pawlak, 1982] Pawlak, Z. (1982). Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356.

[Pawlak, 1996] Pawlak, Z. (1996). Rough sets and data analysis. In *Fuzzy Systems Symposium, 1996. Soft Computing in Intelligent Systems and Information Processing., Proceedings of the 1996 Asian*, pages 1–6. IEEE.

[Pawlak, 2012] Pawlak, Z. (2012). *Rough sets: Theoretical aspects of reasoning about data*, volume 9. Springer Science & Business Media.

[Pedro et al., 2019] Pedro, R. W. D., Machado-Lima, A., and Nunes, F. L. (2019). Is mass classification in mammograms a solved problem?-a critical review over the last 20 years. *Expert Systems with Applications*, 119:90–103.

[Pérez et al., 2015] Pérez, N. P., López, M. A. G., Silva, A., and Ramos, I. (2015). Improving the mann–whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artificial intelligence in medicine*, 63(1):19–31.

[Petrick et al., 1999] Petrick, N., Chan, H.-P., Sahiner, B., and Helvie, M. A. (1999). Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Medical physics*, 26(8):1642–1654.

[Qu et al., 2018] Qu, Y., Shang, C., Mac Parthaláin, N., Wu, W., and Shen, Q. (2018). Multi-functional nearest-neighbour classification. *Soft Computing*, 22(8):2717–2730.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[Riza et al., 2015] Riza, L. S., Bergmeir, C. N., Herrera, F., Benítez Sánchez, J. M., et al. (2015). frbs: Fuzzy rule-based systems for classification and regression in r. American Statistical Association.

[Roffo et al.,] Roffo, G., Melzi, S., and Cristani, M. Feature selection library (matlab toolbox).

[Ross, 2005] Ross, T. J. (2005). *Fuzzy logic with engineering applications*. John Wiley & Sons.

[Rupert Jr et al., 2012] Rupert Jr, G. et al. (2012). *Simultaneous statistical inference*. Springer Science & Business Media.

[Samuels, 1998] Samuels, T. (1998). Illustrated breast imaging reporting and data system birads. *American College of Radiology Publications,*.

[Shaffer, 1995] Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584.

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

[Shen and Boongoen, 2012] Shen, Q. and Boongoen, T. (2012). Fuzzy orders-of-magnitude based link analysis for qualitative alias detection. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):649–664.

[Shen and Chouchoulas, 2002] Shen, Q. and Chouchoulas, A. (2002). A rough-fuzzy approach for generating classification rules. *Pattern Recognition*, 35(11):2425–2438.

[Shen and Yang, 2011] Shen, Q. and Yang, L. (2011). Generalisation of scale and move transformation-based fuzzy interpolation. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 15(3):288–298.

[Shen et al., 1988] Shen, Z., Ding, L., and Mukaidono, M. (1988). Fuzzy resolution principle. In *[1988] Proceedings. The Eighteenth International Symposium on Multiple-Valued Logic*, pages 210–215. IEEE.

[Shen et al., 1993] Shen, Z., Ding, L., and Mukaidono, M. (1993). Methods of revision principle. *Proc. 5th IFSA World Congr*, pages 246–249.

[Shin et al., 2005] Shin, K.-S., Lee, T., and Kim, H.-J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1):127–135.

[Steyerberg, 2008] Steyerberg, E. (2008). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media.

[Sun et al., 2017] Sun, X., Hu, Y., Zhang, L., Chen, Y., Li, P., Xie, Z., and Liu, Z. (2017). Camera-assisted video saliency prediction and its applications. *IEEE Transactions on Cybernetics*.

[Terano et al., 2014] Terano, T., Asai, K., and Sugeno, M. (2014). *Applied fuzzy systems*. Academic Press.

[Tikk, 1999] Tikk, D. (1999). *Investigation of fuzzy rule interpolation and the universal approximator property of fuzzy controllers*. PhD thesis, PhD thesis, Department of Telecommunications and Telematics.

[Tikk and Baranyi, 2000] Tikk, D. and Baranyi, P. (2000). Comprehensive analysis of a new fuzzy rule interpolation method. *IEEE Transactions on Fuzzy Systems*, 8(3):281–296.

[Tikk et al., 1999] Tikk, D., Baranyi, P., Yam, Y., and Koczy, L. T. (1999). Stability of a new interpolation method. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, volume 3, pages 7–9. IEEE.

[Tikk et al., 2011] Tikk, D., Csaba Johanyák, Z., Kovács, S., and Wong, K. W. (2011). Fuzzy rule interpolation and extrapolation techniques: Criteria and evaluation guidelines. *Journal of advanced computational intelligence and intelligent informatics*, 15(3):254–263.

[Tikk et al., 2002] Tikk, D., Joó, I., Kóczy, L., Várlaki, P., Moser, B., and Gedeon, T. D. (2002). Stability of interpolative fuzzy kh controllers. *Fuzzy Sets and Systems*, 125(1):105–119.

[Tikk et al., 1997] Tikk, D., Kóczy, L., and Moser, B. (1997). Stability of interpolative fuzzy kh-controllers. multi-dimensional case. In *Proc. of the Tempus JEP Modify Symposium*, page 8.

[Turksen and Zhong, 1988] Turksen, I. and Zhong, Z. (1988). An approximate analogical reasoning approach based on similarity measures. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(6):1049–1056.

[Ughetto et al., 2000] Ughetto, L., Dubois, D., and Prade, H. (2000). Fuzzy interpolation by convex completion of sparse rule bases. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)*, volume 1, pages 465–470. IEEE.

[Uğuz, 2011] Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, 24(7):1024–1032.

[Vadivel and Surendiran, 2013] Vadivel, A. and Surendiran, B. (2013). A fuzzy rule-based approach for characterization of mammogram masses into bi-rads shape categories. *Computers in biology and medicine*, 43(4):259–267.

[Vass et al., 1992] Vass, G., Kalmár, L., and Kóczy, L. (1992). Extension of the fuzzy rule interpolation method. In *Proc. Int. Conf. Fuzzy Sets Theory Applications*, pages 1–6.

[Vincze and Kovács, 2008] Vincze, D. and Kovács, S. (2008). Using fuzzy rule interpolation based automata for controlling navigation and collision avoidance behaviour of a robot. In *2008 IEEE International Conference on Computational Cybernetics*, pages 79–84. IEEE.

[Wang and Mendel, 1992] Wang, L.-X. and Mendel, J. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man and Cybernetics*, 22(6):1414–1427.

[Wang et al., 2016] Wang, T., Han, Z., Zhao, J., and Wang, W. (2016). Adaptive granulation-based prediction for energy system of steel industry. *IEEE transactions on cybernetics*.

[Wong and Gedeon, 2000] Wong, K. and Gedeon, T. (2000). Petrophysical properties prediction using self-generating fuzzy rules inference system with modified alpha-cut based fuzzy interpolation. In *Proceedings of The Seventh International Conference of Neural Information Processing ICONIP (2000), Korea, (November 2000)*, pages 1088–1092.

[Wong et al., 2000] Wong, K. W., Gedeon, T., and Tikk, D. (2000). An improved multidimensional α-cut based fuzzy interpolation technique. In *International Conference on Artificial Intelligence in Science and Technology (AISAT 2000)*, pages 16–20. University of Tasmania.

[Wong et al., 2005] Wong, K. W., Tikk, D., Gedeon, T. D., and Kóczy, L. T. (2005). Fuzzy rule interpolation for multidimensional input spaces with applications: A case study. *IEEE transactions on fuzzy systems*, 13(6):809–819.

[Wu et al., 2018] Wu, J., Zhou, B., Peck, D., Hsieh, S., Dialani, V., Mackey, L., and Patterson, G. (2018). Deepminer: Discovering interpretable representations for mammogram classification and explanation. *arXiv preprint arXiv:1805.12323*.

[Wu and Schölkopf, 2007] Wu, M. and Schölkopf, B. (2007). A local learning approach for clustering. In *Advances in neural information processing systems*, pages 1529–1536.

[Wu et al., 1996] Wu, Z. Q., Masaharu, M., and Shi, Y. (1996). An improvement to kóczy and hirota's interpolative reasoning in sparse fuzzy rule bases. *International Journal of Approximate Reasoning*, 15(3):185–201.

[Xie et al., 2016] Xie, W., Li, Y., and Ma, Y. (2016). Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*, 173:930–941.

[Xu et al., 2018] Xu, M., Han, M., Qiu, T., and Lin, H. (2018). Hybrid regularized echo state network for multivariate chaotic time series prediction. *IEEE Transactions on Cybernetics*.

[Yam et al., 1999] Yam, Y., Baranyi, P., Tikk, D., and Kóczy, L. (1999). Eliminating the abnormality problem of α-cut based interpolation. In *Proc. 8th IFSA World Congress*, volume 2, pages 726–766.

[Yam and Kóczy, 1998] Yam, Y. and Kóczy, L. (1998). Cartesian representation for fuzzy interpolation. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, volume 3, pages 2936–2937. IEEE.

[Yam and Kóczy, 2001] Yam, Y. and Kóczy, L. (2001). Fuzzy interpolation with cartesian representation and extensibility functions. In *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, pages 2852–2857. IEEE.

[Yam and Kóczy, 2000] Yam, Y. and Kóczy, L. T. (2000). Representing membership functions as points in high-dimensional spaces for fuzzy interpolation and extrapolation. *IEEE Transactions on Fuzzy Systems*, 8(6):761–772.

[Yam et al., 2000a] Yam, Y., Kreinovich, V., and Nguyen, H. T. (2000a). Extracting fuzzy sparse rules by cartesian representation and clustering. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, volume 5, pages 3778–3783. IEEE.

[Yam et al., 2006] Yam, Y., Wong, M., and Baranyi, P. (2006). Interpolation with function space representation of membership functions. *IEEE Transactions on Fuzzy Systems*, 14(3):398–411.

[Yam et al., 2000b] Yam, Y., Wong, M., and Baranyi, P. Z. (2000b). Interpolation as mappings between cartesian spaces. *Pro. Int. Conf. Intelligent Technologies*, 3:459–466.

[Yang et al., 2018] Yang, J., Shang, C., Li, Y., Li, F., and Shen, Q. (2018). Generating anfiss through rule interpolation: An initial investigation. In *UK Workshop on Computational Intelligence*, pages 150–162. Springer.

[Yang et al., 2019] Yang, J., Shang, C., Li, Y., Li, F., and Shen, Q. (2019). Anfis construction with sparse data via group rule interpolation. *IEEE Transactions on Cybernetics*.

[Yang et al., 2017] Yang, L., Chao, F., and Shen, Q. (2017). Generalized adaptive fuzzy rule interpolation. *IEEE Transactions on Fuzzy Systems*, 25(4):839–853.

[Yang and Shen, 2011] Yang, L. and Shen, Q. (2011). Adaptive fuzzy interpolation. *IEEE Transactions on Fuzzy Systems*, 19(6):1107–1126.

[Yang and Shen, 2013] Yang, L. and Shen, Q. (2013). Closed form fuzzy interpolation. *Fuzzy Sets and Systems*, 225:1–22.

[Yassin et al., 2018] Yassin, N. I., Omran, S., El Houby, E. M., and Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, 156:25–45.

[Yi et al., 2017] Yi, D., Sawyer, R. L., Cohn III, D., Dunnmon, J., Lam, C., Xiao, X., and Rubin, D. (2017). Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. *arXiv preprint arXiv:1705.06362*.

[Yuan and Shaw, 1995] Yuan, Y. and Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and systems*, 69(2):125–139.

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

[Zadeh, 1973] Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *Systems, Man and Cybernetics, IEEE Transactions on*, (1):28–44.

[Zadeh, 1994] Zadeh, L. A. (1994). Fuzzy logic, neural networks, and soft computing. *Communications of the ACM*, 37(3):77–85.

[Zeng and Cheung, 2010] Zeng, H. and Cheung, Y.-m. (2010). Feature selection and kernel learning for local learning-based clustering. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1532–1547.

[Zimmermann, 2011] Zimmermann, H.-J. (2011). *Fuzzy set theory—and its applications*. Springer Science & Business Media.